



DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

Dipòsit Legal: T-1809-2011

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tesisenxarxa.net) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tesisenred.net) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tesisenxarxa.net) service has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized neither its spreading and availability from a site foreign to the TDX service. Introducing its content in a window or frame foreign to the TDX service is not authorized (framing). This rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

Determination of banned Sudan dyes in culinary spices through spectroscopic techniques and multivariate analysis

Carolina Vanesa Di Anibal

Doctoral Thesis



ROVIRA I VIRGILI UNIVERSITY

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

Determination of banned Sudan dyes in culinary spices through spectroscopic techniques and multivariate analysis

Carolina Vanesa Di Anibal

DOCTORAL THESIS

Supervisors:

Dr. Itziar Ruisánchez Capelástegui

Prof. María Pilar Callao Lasmarías

Department of Analytical and Organic Chemistry

Universitat Rovira i Virgili

Tarragona 2011

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011



ROVIRA I VIRGILI UNIVERSITY
Department of Analytical Chemistry
and Organic Chemistry

Dr. Itziar Ruisánchez Capelástegui, Associate professor, and Dr. María Pilar Callao Lasmarías, Professor, both of the Department of Analytical Chemistry and Organic Chemistry at Rovira i Virgili University,

CERTIFY:

The doctoral thesis entitled: “DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES AND MULTIVARIATE ANALYSIS”, presented by Carolina Vanesa Di Anibal to receive the degree of Doctor from the Rovira i Virgili University, has been carried out under our supervision in the Chemometrics, Qualimetrics and Nanosensors Group at Rovira i Virgili University, and all results presented in this thesis were obtained in experiments conducted by the mentioned doctoral student.

Tarragona, December 2011

Dr. Itziar Ruisánchez Capelástegui

Prof. María Pilar Callao Lasmarías

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

*El hierro se oxida cuando no se usa, el agua estancada pierde su pureza y se congela
con el frío; asimismo, la inacción absorbe el vigor de la mente.*

Leonardo Da Vinci

Tanto si crees que puedes como si crees que no puedes, estás en lo cierto.

Henry Ford

*Educar a una persona no es hacerle aprender algo que no sabía sino
hacer de él alguien que no existía*

Facundo Geres

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

A mi querido Rodolfo

A mi familia

En memoria de Nicolita y Jorge

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

Recuerdo con mucho cariño el momento en el que se me presento la posibilidad de venir a Tarragona a realizar el doctorado. Una enorme expectación, incertidumbre, entusiasmo, pero sobre todo la certeza de saber que sería una experiencia única en mi vida. Y realmente lo ha sido. Por eso quiero agradecer a todas las personas que han formado parte de este proyecto, como así también a aquellas que han ayudado directa o indirectamente.

En primer lugar quisiera agradecer a mis directoras de tesis, María Pilar Callao e Itziar Ruisánchez. Gracias por haberme proporcionado la oportunidad de llevar a cabo mi tesis doctoral dentro del grupo de Quimiometría, Cualimetría y Nanosensores. Gracias por confiar en mí, por sus consejos, su dedicación, su gran paciencia en todo momento y por los conocimientos brindados que han permitido seguir forjando mi carrera investigadora.

Gracias también a Marisol Larrechi, F. Xavier Rius, Jordi Riu, Joan Ferré, Ricard Boqué, Francisco Andrade y al resto de miembros del grupo de investigación. Ha sido un placer contar con su compañía. He aprendido mucho a nivel científico y personal junto a su lado. Quiero hacer un especial agradecimiento a todas esas personas que constituyen una importante parte del doctorado: mis compañeros doctorandos. Afortunadamente son muchos los que me han acompañado durante el trayecto; algunos están y otros ya no. Todos ustedes han dejado un pedacito de su persona en mí, gracias por haber estado conmigo compartiendo muy gratos y variados momentos y por haberme sacado sonrisas en innumerables momentos. Mi experiencia de formación doctoral contiene muchísimo del aporte de ustedes; por eso les estoy muy agradecida por hacer que la misma haya sido más que gratificante. Todos esos valiosos recuerdos que poseo junto a ustedes durante estos años serán imborrables en mi memoria!

No puedo dejar de nombrar a Santiago Macho. Santi te agradezco mucho por haberme introducido con Matlab, por el tiempo y paciencia depositados en mí y por haber estado siempre con una gran disposición para atender mis consultas.

Gracias también a esas personas que amablemente han estado siempre dispuestas a cualquier cosa que necesitara. Gracias a Marta Odena, Eulalia, Jaume, Tere, Olga y Avelina.

Quisiera agradecer al personal del SRCiT de la URV por la constante ayuda y dedicación que han tenido conmigo. Gracias a Ramón Guerrero, Miguel Ángel Rodríguez, Mercè Moncusí y Pablo Ramos.

Gracias al resto de personas que están al otro lado, a muchos kilómetros de distancia pero siempre cerca de mí. A mi gran familia que ha constituido mi espalda durante todo este trayecto. Esta tesis esta dedicada a cada uno de ustedes y no hubiera tenido el mismo sentimiento sin el valioso aporte que le han conferido a lo largo de estos años.

A mamá y papá. Se lo mucho que les reconforta la realización de este logro pero quiero que sepan que no podría haberlo obtenido sin ustedes. Infinitas gracias por el incondicional apoyo que me han dado, el empuje a veces casi vital y sobre todo por haber confiado en mí en todo momento. Nunca han dudado de mi capacidad para finalizar todo lo que comienzo y eso ha sido un gran aliciente en todo este tiempo. Gracias por enseñarme a poner mucho amor en cada cosa que emprendo en la vida, por haberme inculcado con mucho cariño esos valores que han permitido que hoy llegara a donde estoy. Mamá y papá, se la felicidad que les proporciona este logro y en gran parte es para agradecerles la inmensa cantidad de cosas buenas que me han proporcionado en la vida. A pesar de la distancia, mi corazón siempre ha estado y estará junto a ustedes.

A mis queridos hermanos: Vero, Nico, Lau y Eva. Mis adorados hermanitos del alma, mis amigos, ustedes saben el gran lugar que ocupan en mi corazón y lo muchísimo que los quiero. Gracias por haberme hablado sinceramente siempre con el corazón, por haberme estado acompañando en cada instante (sea bueno o no tan bueno) de este doctorado, por haberme hecho sentir siempre tan cerca de casa, por la alegría que

me han transmitido en cada momento. Su compañía durante todo este tiempo ha sido esencial para mí, chicos sin ustedes esta experiencia no hubiera sido la misma. Gracias por el respaldo que me han dado y por demostrarme que el contar con el amor que me han proporcionado durante todo este tiempo ha sido valiosísimo para enfrentar cualquier situación. Mi felicidad está íntimamente relacionada con la de ustedes.

A Fausta y Néstor. Gracias por ser dos maravillosas personas que siempre me han recordado que puedo seguir adelante a pesar de los pequeños obstáculos que pone el camino de la vida. Gracias por haber acortado las distancias de múltiples maneras que han hecho que muchas sonrisas nazcan en oportunos momentos. ¿Que son varios miles de kilómetros? ¡Pues nada! Les agradezco profundamente por todo el amor, el cariño, el afecto y la confianza que me han aportado a lo largo de este trayecto. La presencia de estos sentimientos han sido más que importantes para mí, incluso esenciales en muchos momentos. Se el gran significado que representa la obtención de este logro en sus vidas, este forma parte de ustedes.

Al resto de mi gran familia. Por suerte son muchos y quiero agradecerles a todos ustedes por formar parte de esta experiencia: A los abuelos, tíos y primos que siempre han estado presente en este proyecto y han depositado en mí una gran confianza a lo largo de estos años de distancia (que nunca la he sentido como tal porque siempre han estado acompañándome). Me enorgullece el hecho de poder contar con el cálido apoyo y cariño que siempre me dan, gracias por ser una estupenda familia! Gracias por la alegría, las risas y todos los consejos dados en los momentos oportunos. Gracias por la incondicional presencia y por el continuo aliento para seguir delante. Ustedes saben lo mucho que los quiero y aprecio a todos ustedes! Gracias a las abuelas Carmen y Laura, al abu Pascal, a Norma, Martín y Gastón. Gracias familia Pascal, Tumminello, Concellón, Papagni, Geres y Di Anibal.

No puedo dejar de agradecer a mis amigos que siempre han estado apoyándome con su presencia, su afecto y ese gran empuje que siempre me han dado. Gracias Vani,

Gise, Fer, Edu, Marce, Belén, Nico, Lauri y Ernesto. Gracias por haberme mostrado el verdadero sentimiento de amistad. Esto no hubiera tenido el mismo encanto sin la presencia de ustedes, por eso el aporte que me han hecho es muy valioso para mí.

Por último, pero no por eso menos importante, infinitas gracias a Rodolfo, mi compañero de vida. Rodó te animaste a iniciar conmigo este viaje hacia lo desconocido y te estoy profundamente agradecida por ello, ya que junto a vos sume ganas, entusiasmo, valor y confianza para abordar esta experiencia. No dudaste nunca en mis capacidades, me alentaste en todo momento y tu ayuda, compañía, entrega, compañerismo, dedicación y sobre todo tu amor han logrado que esta etapa finalice de la mejor manera posible. Gracias por proporcionarme el respaldo y la guía necesarios para que este logro se concrete. Tengo el privilegio de poder compartir la alegría de este momento con vos, este gran momento que representa nuestra felicidad. Este periodo que culmina no tendría esa mágica esencia sin la huella de tu presencia y esa gran inspiración que me has proporcionado. Gracias, desde lo más profundo de mi corazón. Las palabras no me alcanzan para agradecerte todo lo que has hecho.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION AND OBJECTIVES	1
1.1. SCOPE	3
1.2. OBJECTIVES	5
1.3. STRUCTURE	5
1.4. REFERENCES	8
CHAPTER 2. THEORETICAL AND PRACTICAL ASPECTS	9
2.1. ANALYTES	11
2.2. SAMPLES	13
2.3. ANALYTICAL TECHNIQUES	14
2.3.1. UV-Visible Spectroscopy	16
2.3.2. Proton Nuclear Magnetic Resonance Spectroscopy	17
2.3.3. Raman Spectroscopy	19
2.4. MULTIVARIATE CHEMOMETRICS TOOLS	21
2.4.1. Principal Components Analysis (PCA)	22
2.4.2. Classification techniques	23
2.4.2.1. <i>K</i> -Nearest Neighbours (<i>KNN</i>)	26
2.4.2.2. Soft Independent Modelling of Class Analogy (SIMCA)	27
2.4.2.3. Partial Least Squares-Discriminant Analysis (PLS-DA)	28

2.4.3. Variable Selection Techniques	29
2.4.4. Data Fusion	31
2.4.5. Multivariate Standardization	32
2.5. REFERENCES	36
CHAPTER 3. EXPERIMENTAL PART AND RESULTS	43
3.1. SPECTROSCOPIC TECHNIQUES AND MULTIVARIATE ANALYSIS FOR THE DETERMINATION OF SUDAN I TO IV DYES IN CULINARY SPICES	45
3.1.1. Determining the adulteration of spices with Sudan I-II-III-IV dyes by UV-Visible spectroscopy and multivariate classification techniques	47
3.1.2. High-Resolution ¹ H-Nuclear Magnetic Resonance spectrometry combined with chemometrics treatment to identify adulteration of culinary spices with Sudan dyes	71
3.1.3. Raman Spectroscopy and chemometrics as a tool for detecting Sudan I dye in culinary spices	93
3.2. IMPLEMENTATION OF CHEMOMETRIC STRATEGIES TO IMPROVE THE CLASSIFICATION RESULTS	117
3.2.1. ¹ H-NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs	119
3.2.2. ¹ H-NMR and UV-Visible data fusion for determining Sudan dyes in culinary spices	143

3.2.3 Standardization of UV-Visible data in a food adulteration classification problem.....	165
CHAPTER 4. GENERAL CONCLUSIONS	189
APPENDIX	195
A.1 PAPER PRESENTED	197
A.2 MEETING CONTRIBUTIONS	198

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

C HAPTER 1

Introduction and Objectives

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

1.1. Scope

In analytical chemistry, screening methods have constituted an effective and important tool over the years in many fields and routine laboratories. From a practical point of view, it is quite usual to make an initial analysis to determine whether the analytes of interest are present or absent in a sample at a certain level and in some cases, if they are present, there may be interest in quantifying these levels. Screening methods have the advantage of reducing both the time and cost of analyses and of allowing timely decisions to be made, usually regarding the use of simpler and less expensive analytical instrumentation. As such, screening methods provide a useful solution in situations requiring a rapid response.

Screening methods have been widely used in environmental, pharmaceutical, medical, clinical and food analysis, and can often be adapted to specific problems. There are many criteria for classifying screening methods [1] and a particular case is constituted by methods based on an unspecific response which might be subjected to be a discriminatory response. In this sense, spectroscopic techniques coupled with multivariate classification have played an important role in analytical chemistry in recent years, and have been the subject of research in different fields. Nevertheless, many scientific aspects regarding the optimal implementation of these groups of techniques still require further research. Classification techniques are used for grouping and classifying objects. For example, they can be used to identify variations in the composition of an adulterated food or to identify changes caused by an industrial process. In general terms, the aim of any classifier is to build a classification rule and use it to assign future unlabeled samples to one of the predefined classes.

Among the multiple applications that the aforementioned methods have, those related to health protection are especially important to the scientific community. In particular, food adulteration and contamination issues have been widely studied. Food adulteration is the act of intentionally adding a prohibited substance to a food to change its chemical composition and/or its physical appearance, and is generally carried out for economic reasons. Because this type of act can have a significant impact on public health, adulterated food products need to be properly checked. This in turn forces government institutions to establish certain commissions and laws to prevent this practice. Specifically, the intentional addition of prohibited Sudan I, II, III and IV dyes to culinary spices, with the aim of intensifying and maintaining their natural appearance, has received increasing attention in the European Union since 2003 [2], where it was reported the first alarm concerning the entrance of adulterated food products into the food chain. Consequently, the EU has adopted regulatory measurements against the use of Sudan dyes in food, calling on member states to organize testing food products to monitor the presence or absence of these adulterants [3].

Over the last few years, a wide range of analytical methodologies has been developed to detect and quantify the use of these dyes as food additives. Most of these methodologies are based on liquid chromatography with different detection systems. Nowadays it is not common to find spices adulterated with Sudan dyes for general sale thanks to all the efforts and the strictly controls that have been implemented. Nevertheless, screening methods seems to be a good means of carrying out routine analysis, and have stimulated the interest of public health regulatory laboratories; such as the Public Health Laboratory of the Catalan Government's Department of Health in Tarragona (Laboratorio de Salud Pública, Departamento de Salud, Generalitat de Catalunya) with whom we are in contact.

1.2. Objectives of the thesis

The main objective of this thesis is to develop multivariate analytical screening methodologies based on spectrometric techniques and chemometric data analysis for determining Sudan I, II, III and IV dyes in culinary spices. This has been achieved by studying, establishing and/or applying the following specific items:

- 1) Spectrometric techniques such as UV-Visible, $^1\text{H-NMR}$ and Raman to obtain the multivariate data (spectra).
- 2) Chemometric tools such as exploratory data analysis, supervised classification techniques, data processing and variable selection techniques to extract the maximum possible information from the spectral data.
- 3) Data fusion strategies to improve the classification of results derived from the individual spectrometric techniques.
- 4) Multivariate transfer techniques (standardization) to maintain the original classification model through a lapse of time.

1.3. Structure

This thesis is structured in four chapters.

Chapter 1: Introduction and objectives. This chapter introduces the framework that will be used to study and implement multivariate screening methods aimed at identifying adulterated foods. The chapter then defines the aims and structure of the thesis.

Chapter 2: Practical and theoretical aspects. This chapter is divided into three parts. The first part describes the different samples (culinary spices) and analytes to be studied (Sudan I, II, III and IV dyes). The second part describes certain theoretical aspects regarding the instrumental analytical techniques used (UV-Visible, NMR and Raman), and shows why these techniques are useful for this kind of analysis. Finally, the third part contains a brief description of the chemometric tools used for data visualization and classification, the variable selection and data processing techniques, and finally the data fusion and multivariate transfer (standardization) methodologies.

Chapter 3: Experimental part and results: This chapter contains the results of the experimental work that has been carried out, which have also been presented in different scientific publications while the thesis was in progress. These publications describe the methods used, the results, and the discussions and conclusions that can be drawn. All this information is sub-divided into two parts. The first part contains three papers regarding the different spectrometric techniques used to detect Sudan dyes in food matrices; and the multivariate techniques used both for classification and exploratory data analysis. Also, the proper chemometric treatment required for each analytical signal is evaluated; these being a variable selection technique in the case of NMR spectra and denoising and background removal in the case of Raman spectra. The second part contains three more papers in which various chemometric approaches were evaluated for the following reasons: to compare the application of three variables selection techniques to NMR data, to fuse the UV-Visible data with NMR data by means of different strategies, and to implement a multivariate standardization technique to evaluate the transfer of classification methods to UV-Visible data measured under different conditions.

Chapter 4: General conclusions: This chapter contains the general conclusions of the thesis.

The last part of the thesis (**Appendix**) contains a list of the papers and meeting presentations attended during this period.

1.3. References

[1] R. Muñoz-Olivas, *Trends in Anal. Chem.* **23** (2004) 203.

[2] FSA (Food Standards Agency), available at:

<http://www.food.gov.uk/foodindustry/guidancenotes/foodguid/sudanguidance>

[3] *Official Journal of the European Union* (2005/402/EC) L135/34.

C HAPTER 2

Practical and Theoretical Aspects

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

2.1. Analytes

Sudan I, II, III and IV dyes are a family of red dyes which are made up of the phenyl, azo and naphthol groups (Figure 2.1). The chemical structures of Sudan II and IV dyes also contain methyl groups. This family of azo-dyes has been mainly used in industry as colorant in various products, and because of their low price they have also unfortunately been used as additives in certain foodstuffs such as culinary spices.

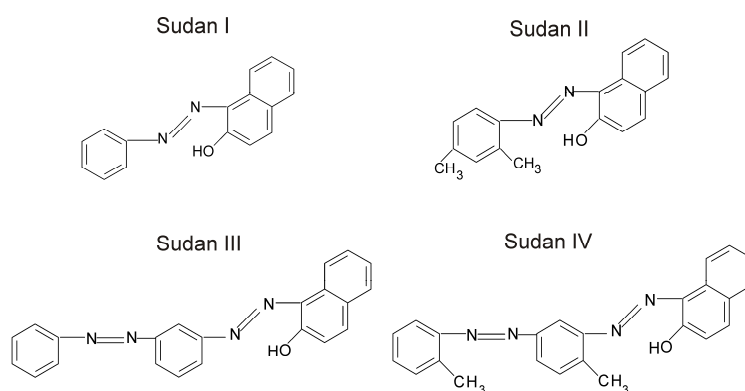


Figure 2.1. Chemical structure of the Sudan I, II, III and IV dyes.

Azo compounds are by far the most widely used synthetic organic colorants; and more than 2000 of these substances are listed in the colour index (CI) [1]. Among the organic colorants, most azo-dyes are recognized as being carcinogens [2] and the genetic toxicity of some of them has been assessed [3].

According to the annual reports of the Rapid Alert System for Food and Feed (RASFF) [4], notifications of foods contaminated with Sudan dyes in Europe, the Near East and Africa reached their highest between 2003 and 2006. Figure 2.2 shows the number of notifications for each year. These adulterations include the Sudan I to IV dyes, Para-Red Sudan dye and some mixtures of these Sudan dyes in

certain culinary spices and processed products containing these spices, sauces and palm oil. The figure shows that the highest number of notifications occurred in 2004 and that the number of cases in 2006 significantly decreased by more than 75%. In 2007 and 2008 the RASFF reported a small decrease in the presence of Sudan dyes and the annual report from 2009 only mentions the cases encountered in previous years.

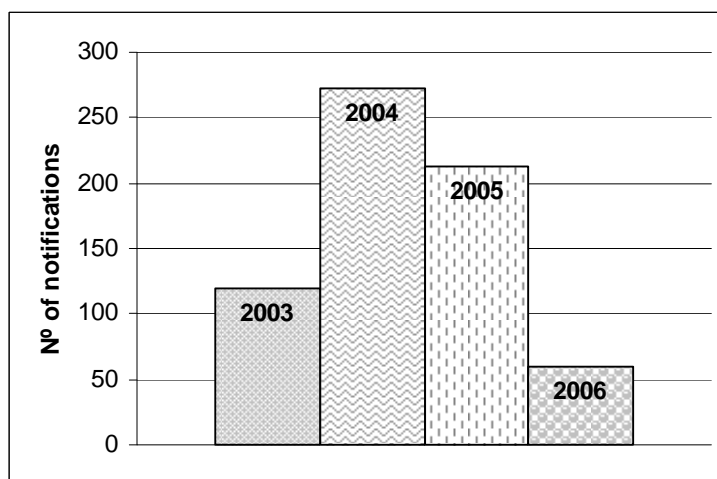


Figure 2.2. Number of notifications of foods adulterated with Sudan dyes found in different countries during the period 2003-2006.

Although the presence of Sudan dyes in foods has decreased in the last few years, they are still used to adulterate certain loose/non-branded food products that are sold mainly in rural and urban markets, and such illegal activities are expected to continue [5]. This situation encourages the development of simple, fast and reliable methods to be used as monitoring tools for detecting Sudan red dyes in food products.

2.2. Samples

Culinary spices are largely used by the global food industry primarily for food processing and culinary preparations with the aim adding aroma, flavour and colour and of preserving food. Colouring food is important to food manufacturers because it allows the desired aesthetic quality to be obtained. Highly coloured spices play an important role in achieving this, but they usually lose their colour over time and thus may become less appealing to consumers. Unfortunately, this has led to the illegal and fraudulent adulteration of culinary spices to make them appear more attractive than they are. Thus, profit is increased by the addition of a cheap compound that is easily disguised in the spice [6].

Among the various types of natural plant that are available for use in culinary preparations, the *capsicum* genus is particularly important in that it provides spices such as chilli powder, hot and mild paprika and cayenne pepper. These products have been subjected to adulteration by Sudan dyes. In 2003 the first European emergency measure was implemented regarding hot chilli powder and related products that were suspected to have been adulterated with the Sudan I dye [7]. Subsequently in 2004, this measure was extended to other Sudan dyes (II, III and IV) and also to curry powder and to all forms of the genus *Capsicum* including hot and mild paprika [8]. Finally, other foods such as turmeric and palm oil were included in 2005 [9].

This thesis focuses particularly on the use of Sudan dyes in turmeric, curry and mild and hot paprika. However, Sudan dyes not only been used to adulterate culinary spices, but have also been used in other food products. Some examples are summarized in Table 2.1.

Table 2.1. Sudan I-IV and Sudan related-dyes found in different food products.

Sudan dyes	Related dyes	Food product	Reference
I-IV	Red B and Red 7b	chilli and sausage sauce / egg yolk	10
I, III	Para red	chilli sauce	11
I-IV		chilli sauce including tomato, vinegar and meat sauce	12
I-IV		tomato sauce	13
	Para Red	pickles / tabasco sauce	14
IV		duck eggs	15
I-IV		eggs	16
I-IV	Red G, Red 7b and Para red	duck muscle / eggs	17
I-IV		ketchup / eggs	18
I-IV	Red G, Red 7b and Para red	duck meat / pepper sauce	19
I-IV		sausages	20
I-IV		palm oil	21
I-IV		soft drinks	22
I		pepper paste / chorizo	23
I		oven-baked food	24

2.3. Analytical Techniques

Over the years, there has been an increasing interest in developing methodologies for monitoring analytical parameters in real time. Specifically in the field of food safety, the desire has been to carry out product analyses in a simple, rapid and qualitative or quantitative manner.

Many different methods have been proposed to determine the presence of the Sudan I to IV dyes in foods. The most commonly used methods have been

those based on liquid chromatography (LC) with different detection systems, and recent advances in certain pre-extraction methods have also been reported [12,15,20]. A detailed review of the different analytical techniques used for determining these dyes in food matrices is given by Rebane R. et al [25]. There have also been advances in the enzyme-linked immunosorbent assay (ELISA) techniques [11,26,27]. However, in general most of these techniques are time consuming; and require expensive instrumentation and clean-up and pre-concentration processes to achieve satisfactory results.

This thesis focuses on the use of three spectroscopic techniques: UV-Visible, $^1\text{H-NMR}$ and Raman. These techniques have the advantage of providing an analytical response quickly, and the significant progress they have had in recent years includes the possibility for using with portable equipments so that on-site analysis can be carried out. These developments have gained a great deal of attention because of their time and cost benefits. However, the techniques still lack specificity when working with complex matrices such as foods. Because of this, it is important to use an appropriate multivariate signal treatment to obtain useful information from spectra, and within this context chemometrics tools provide a wide range of algorithms depending on the nature of the problem.

The UV-Visible, $^1\text{H-NMR}$ and Raman techniques were selected because they provide rich and valuable molecular information concerning both the sample and the analytes under study. This enables us to evaluate them by: 1) applying the techniques individually and 2) applying the techniques together to determine the synergism obtained from complementary information. This situation opens the way to addressing the adulteration problem from different perspectives.

2.3.1. UV-Visible

UV-Visible spectroscopy is one of the most widely used techniques in laboratories that are concerned with identifying and measuring organic and inorganic compounds in a wide range of products and industrial processes. Particularly in the area of food, a UV-Visible spectrophotometer has become an essential tool for both research and routine control analysis, because it is quick, simple, inexpensive and robust.

UV-Visible spectroscopy is a well established technique that measures a sample's absorption of electromagnetic radiation in the spectral region between approximately 180 and 700 nm. Figure 2.3 presents a scheme showing the instrumental configuration of a UV-Visible spectrophotometer with a photodiode array detector (DAD).

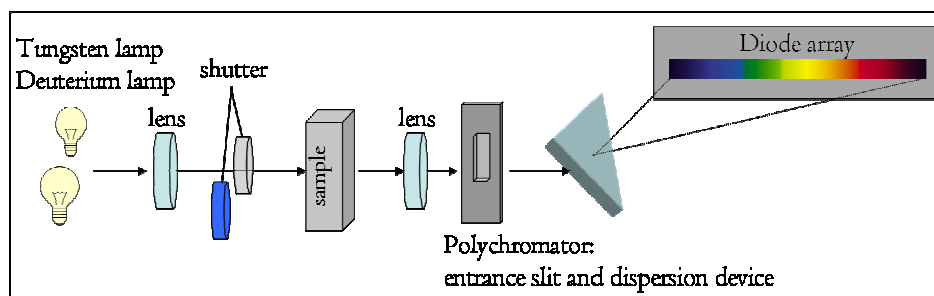


Figure 2.3. Scheme of a UV-Visible-DAD spectrometer

When UV-Visible spectroscopy has been applied to food products, it has been widely used as detector in chromatographic techniques, but it has barely been evaluated as a single analytical technique. Table 2.2 summarises some examples of

the application of UV-Visible spectroscopy as a single analytical technique in foods.

Table 2.2. Examples of different applications of UV-Visible spectroscopy in foods.

Sample	Analytes	Study	Reference
Turmeric, curry and mild/hot paprika	Sudan I-IV dyes	Adulteration	28
Turmeric	Turmerone	Flavouring and biological properties	29
Saffron	Safranal	Quality control	30
Milk	Lipids	Fat determination	31
Tomato fruit and juice	Lycopene	Quantification	32
Sausages	<i>Lactobacillus fermentum</i>	Curing meat process	33
Fruits	Ascorbic Acid	Quantification	34
Honey	Hydroxymethylfurfural (HMF)	Quantification	35
Food supplement	Dihydromyricetin–Lecithin	Properties and antioxidant activity	36
Thermally treated meat	Maillard-derived compounds	Maillard reaction	37

2.3.2. Nuclear Magnetic Resonance (NMR)

Proton Nuclear Magnetic Resonance spectroscopy ($^1\text{H-NMR}$) is a powerful analytical tool traditionally used to elucidate the structure of compounds within the field of organic chemistry. Henceforward, we will refer to proton NMR spectroscopy as NMR. This technique is based on studying the spinning mechanism of hydrogen nuclei present in the molecules when exposed to a strong magnetic field. NMR provides a wealth of detailed information about the molecules and their environment. Consequently, an NMR spectrum can be considered to be a “fingerprint” of the sample under study. Figure 2.4 shows the scheme of a classic NMR instrument.

In recent years, improvements in the sensitivity and resolution of NMR have meant that it has been increasingly used for solving food problems [38]. Such advances have motivated scientists to find new applications of this technique within the field of food. In the context of food quality, NMR has been employed to detect adulteration and to authenticate and control the manufacture and properties of food products. Table 2.3 gives some examples of NMR with different modalities and advances in instrumentation applied to food analysis.

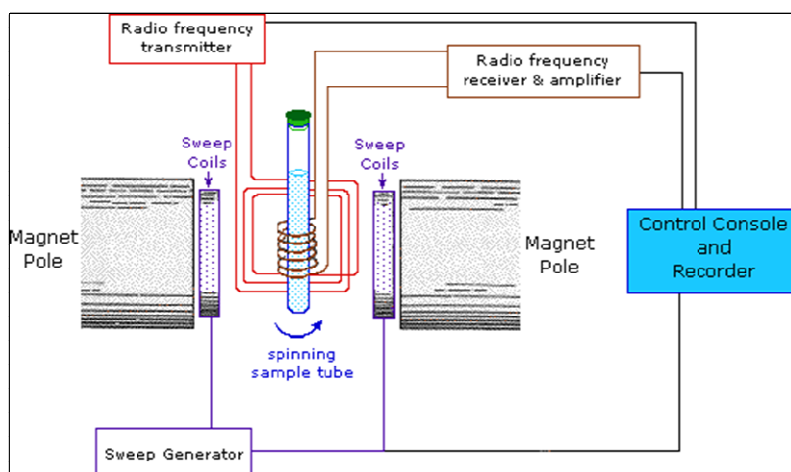


Figure 2.4. Scheme of an NMR instrument.

Table 2.3. NMR with different modalities and advances in instrumentation applied to food analysis.

NMR modality	Sample	Study	Reference
High resolution NMR	Culinary spices	Adulteration	39
Low-field NMR	Salmon and other foods	Food properties	40,41
Flow-injection NMR	Fruit juice	Quality control	42
Portable NMR probe	Food fluids	Food properties	43
Site-specific natural isotope fractionation (SNIF) NMR	Fruit juice	Exogenous compounds	44
Two dimensional (2D)-NMR	Tomato juice	Characterization of carotenoids	45
Magnetic resonance image (MRI)	Coconut	Post harvesting maturing	46
Pulse field gradient (PFG) NMR	Butter	Emulsion properties	47
Solid state-NMR	Foods	Structure and dynamics of food compounds	48

2.3.3. Raman spectroscopy

Raman spectroscopy is a vibrational technique that has made remarkable progress over the last few years owing to technological innovations in lasers, spectrometers and detectors. These developments have made Raman spectroscopy a viable and useful analytical technique in chemistry, and one which continues to grow in popularity in analytical laboratories. The Raman effect basically refers to the shifts in the wavelength or frequency of an incident beam of radiation that are caused by inelastic scattering on an interaction between the photons and the sample. A Raman spectrum contains a wealth of rich information about the sample under study. Figure 2.5 shows the scheme of a Raman confocal microscope with a Charge Coupled Device (CCD) camera used as Raman detector.

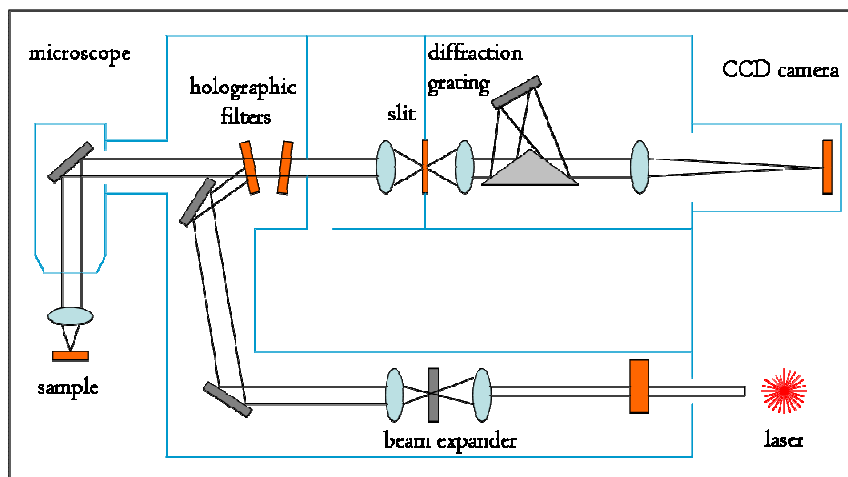


Figure 2.5. Scheme of a Raman microscope

This technique can be applied in many analytical contexts and is versatile. It can perform measurements under multiple modalities, for example, Surface-Enhanced Raman has increased sensitivity, Raman microspectroscopy improves spatial resolution and Resonance Raman obtains specific information. Today, although Raman spectroscopy is used in many areas of science and technologies, it is still an emerging technique in food analysis, despite the considerable promise it offers in the identification, structural investigation and qualitative and quantitative analysis of constituents, additives and contaminations in various food matrices [49]. Table 2.4 cites some examples of the application of Raman spectroscopy in different food ambits. Nonetheless, the literature concerning the study of Sudan dyes with Raman spectroscopy is very limited; therefore there is plenty of scope for research into this technique's suitability for determining the presence of these banned dyes in food matrices.

Table 2.4. Examples of Raman applications in different food ambits.

Raman modality	Sample	Study	Reference
SERS	Milk	Adulteration	50
Micro-Raman	Milk	Characterization of fat globules	51
SERS	Foods	Microbiological contamination	52
Conventional/FT-Raman and micro-Raman	Muscle food	Protein structure	53
Portable Raman spectrometer	Olive oil	Authentication	54
FT-Raman	Fructose and honey	Chemical changes in carbohydrates	55
Conventional Raman	Pork meat	<i>Structural changes in:</i> food ageing and cooking	56
Conventional Raman	Fish	food freezing	57
FT-Raman	Sunflower and mustard leaves	Interactions between plants	58
FT-Raman	Edible oils	Lipids oxidation	59
FT-Raman	Rice	Rice globulin conformation	60

2.4. Multivariate Chemometric Tools

Chemometrics is a well known science to analytical chemists who use mathematical and statistical methods to obtain relevant information by analyzing chemical data [61]. Most chemical measurements are inherently multivariate, because more than one value can be obtained from a single sample, as in the case of a spectroscopic analysis in which the spectrum of each sample can be recorded in a hundreds of wavelengths. Multivariate analysis plays an important role in chemometrics because it provides a means for adequately treating such data.

The large amount of data produced by modern instrumentation together with the availability of powerful computational technology means that multivariate data analysis has become an essential tool in analytical chemistry for extracting the maximum useful information from data matrices.

2.4.1. Principal Component Analysis (PCA)

PCA looks for a linear combination of the original variables, so it decomposes the original matrix X containing the chemical data by a product of two matrices: the score and the loading matrices according to:

$$\mathbf{X} = \mathbf{T} \mathbf{L}^T \quad \text{Eq. (2.1)}$$

where \mathbf{X} consists of n rows (usually samples) and p columns (usually variables); \mathbf{T} is the scores matrix with n rows and d columns (number of principal components) and \mathbf{L} is the loadings matrix with d columns and p rows. Each column in the scores matrix is a transformed variable which it is built by multiplying the weight or influence (loadings matrix) that each variable has in the PCA model by each original variable (2.2).

$$\mathbf{T} = \mathbf{X} \mathbf{L} \quad \text{Eq. (2.2)}$$

The scores matrix contain information about the samples which are described in terms of their projection onto the PCs, and the loading matrix contain information about the variables that indicates which ones are the most important for describing variation in the original data.

The principal components (PCs) are determined on the basis of the maximum variance criterion. The first principal component (t_1) explains the maximum amount of variance in the data; and each subsequent principal component (t_2, \dots, t_d) describes a decreasing amount of variance that is not modelled by the previous components (Figure 2.6). The PCs constitute a way to project high dimensional data onto a lower dimensional space.

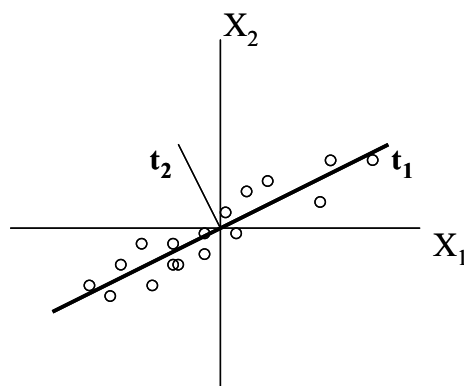


Figure 2.6. PCA score plot showing a systematic pattern of variation in a data set.

PCA can be used for different purposes such as data visualization technique (exploratory analysis), data reduction technique and in some classification and calibration techniques, among others. One of PCA's main uses has been to get an initial idea about the inherent behaviour of the data to be analysed. PCA is useful for obtaining a graphical visualization on a scores plot to show how information is spanned into the space defined by the first PCs. It is also useful for determining relations and trends in the data.

2.4.2. Classification Techniques

A pattern recognition method can be described as a supervised approach if there is a *priori* knowledge about the grouping in the data. The classification process uses known objects belonging to the different classes to fit a multivariate model or to establish a decision rule, which is further used to assign unknown objects to one, more than one or none of the existing classes.

Generally, it is not possible to decide which classification technique will give the best results in advance, since the classification performance depends greatly on the characteristics and nature of the data to be classified, and there is no classifier that works best on all given analytical problems. Therefore, of the various classification techniques that are available, we have implemented the following three ones: *K*-Nearest Neighbours (*KNN*), Soft Independent Modelling of Class Analogy (*SIMCA*) and Partial Least Squares-Discriminant Analysis (*PLS-DA*). The two first techniques are based discriminating and modelling respectively, while the last combines both these properties.

SIMCA and *PLS-DA* are techniques based on factor analysis, i.e., they are characterized by principal components or latent variables. The main difference between these two techniques is that the first focuses on each class independently of the other classes whereas the other one takes into account all the classes to maximize the separation among them. The residuals in *SIMCA* and the data in *PLS-DA* are assumed to be distributed normally, in contrast to *KNN*, which is a simple method that does not make statistical assumptions about the data.

All three techniques (*KNN*, *SIMCA* and *PLS-DA*) have been widely used to solve problems commonly found in foods and natural products. Table 2.5 cites some examples from the last six years.

Table 2.5. Recent examples of some *KNN*, *SIMCA* and *PLS-DA* applications regarding foods and natural products.

Sample	Classification techniques	Classification criteria	Reference
Walnut		Parts of wallnuts	62
Edible oils		Quality and geographical origin	63
Milk, fruit juice and tonic		Brand	64
Wine	<i>KNN</i>	Grape cultivar	65
Apples		Apple variety	66
Citrus fruit		Fruits flaws	67
Honey		Geographical origin	68
Wine			69
Cider	<i>KNN-SIMCA</i>	Geographical origin	70
Fish		Production method and geographical origin	71
Meat	<i>SIMCA</i>	Adulteration	72
Milk		Storage treatment	73
Beer		Brand	74
Honey		Geographical origin	75
Ginseng	<i>SIMCA</i> and <i>PLS-DA</i>	Counterfeit and adulterated	76
Green soybean		Defective pods	77
Rice wine		Ageing time and brand	78
Coffee		Roasting degree	79
Chinese medical plant		Geographical origin	80
Meat		Illicit meat treatment	81
Wheat flour	<i>PLS-DA</i>	Adulteration	82
Extra-virgin olive oil		Adulteration	83
Honey		Adulteration	84
Milk		Adulteration	85

2.4.2.1. K -Nearest Neighbours (KNN)

KNN is a classification technique that calculates the distances between each sample and the rest of the existing samples. The philosophy of this technique is that a sample will be assigned to the class to which the k nearest samples belong, these being the samples positioned at the lowest distance in the multidimensional space. The first stage is to determine the number of neighbours to be used (k); this can be optimized by selecting the k value that gives the lowest prediction error. Then, to classify an unknown sample, KNN selects the k -nearest objects to this sample and a majority rule is applied whereby the unknown sample is classified in the group to which the majority k samples belong (based on the minimum distance). This situation is represented in Figure 2.7 where $k=3$ is used.

This technique has the advantage of being conceptually and mathematically simple, however it does have some limitations: it does not take into account the spread or variance within each class; and the classification should depend on the number of samples contained in each class (which should be approximately equal), otherwise, the sample will be assigned to the class with most representatives.

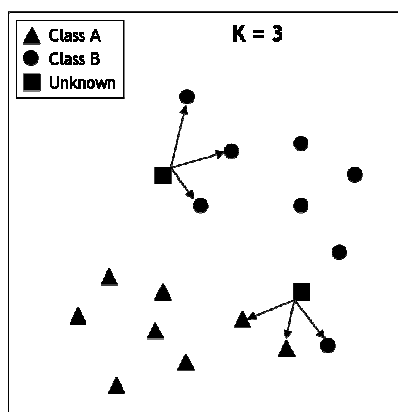


Figure 2.7. Classification rule for the KNN technique.

2.4.2.2. Soft Independent Modelling of Class Analogy (SIMCA)

This classification technique is based on the use of Principal Component Analysis to model each class independently. The first step is to establish the model; thus the so-called SIMCA boxes are constructed for each class by determining the significant number of principal components and by defining the boundary regions (edges of the boxes) around each PCA model (see Figure 2.8).

An unknown sample is first classified into a certain class by means of a statistical F-test, where a comparison is made between the variance of the sample residual in the space defined by the principal components and the residuals of the class model. Furthermore, it is verified whether the unknown falls within the boundaries established for that class. SIMCA provides three possible classification outcomes: samples that belong to a certain class, samples that belong to more than one class and samples that do not belong to any class. In addition, SIMCA offers information about the importance of each variable by describing each class (modelling power) and its discriminant power.

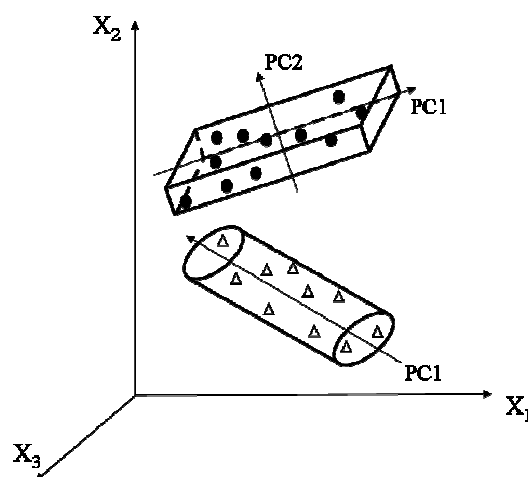


Figure 2.8. SIMCA models for different numbers of significant PCs.

2.4.2.3. Partial Least Squares-Discriminant Analysis (PLS-DA)

Although Partial Least Squares (PLS) was developed as a regression technique, it can also be used for discriminant analysis in the form of PLS-DA. The PLS-DA algorithm establishes a model based on certain factors or latent variables (LVs), which are analogous to the principal components in PCA.

The first step in PLS-DA is to build a model that uses an adequate number of factors and where the X block (original data) and the Y block (classes) are decomposed according to Eqs. 2.3 and 2.3:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad \text{Eq. (2.3)}$$

$$\mathbf{Y} = \mathbf{UQ} + \mathbf{F} \quad \text{Eq. (2.4)}$$

where \mathbf{T} and \mathbf{U} are the scores matrices, and \mathbf{P} and \mathbf{Q} are the corresponding loadings matrices. The error is contained in the \mathbf{E} and \mathbf{F} corresponding residual matrices. Table 2.6 shows an example of class codification containing five classes.

Table 2.6. Identity matrix for classification

		X block:		Y block:		
		1	2	3	4	5
Class 1		1	0	0	0	0
Class 2		0	1	0	0	0
Class 3	Spectral data	0	0	1	0	0
Class 4		0	0	0	1	0
Class 5		0	0	0	0	1

An inner relationship is constructed that relates \mathbf{T} and \mathbf{U} , and once the PLS model is calculated, a \mathbf{b} vector containing the regression coefficients of the model is obtained. The prediction for a sample \mathbf{x} can be obtained according to:

$$y = \mathbf{x}^T \mathbf{b} \quad \text{Eq. (2.5)}$$

The prediction values for each class range from 0 (not belonging) to 1 (belonging), so it is necessary to establish a threshold value between these two values. This threshold is calculated by assuming that the y predicted values follow a Gaussian distribution, which is estimated by using the mean and standard deviation of the y predicted values for each class. PLS-DA was applied in this thesis to solve a multiclass classification problem.

2.4.3. Variable Selection Techniques

When multivariate analysis involves large datasets, variable selection processes play an important role because they eliminate the less significant or non-informative variables. The overall aim of any variable selection technique is to capture variables from the original dataset that are most specifically related to the problem of interest; and to exclude those variables that are affected by other sources of variation and that worsen the multivariate model. When using variables for classification, it is preferable to retain those variables that have high discrimination ability.

Selecting only a subset of the variables rather than the whole set of variables before implementing a classification analysis provides several advantages, such as [86]:

- Improving the model performance (prediction ability).

- Preventing overfitting
- Providing more robust models
- Gaining a better understanding and interpretation of the data.

In literature, many algorithms based on quite different criteria have been proposed for selecting variables. Since no variable selection method provides optimal results for all situations, it is necessary to test several methods to ensure that classification ability is as close as possible to when all the variables are used. Some examples of variable selection methods and their applications in different areas are cited in Table 2.7.

Table 2.7. Examples of some variable selection methods applied in different areas.

Method	Area	Technique	Reference
Fisher's criteria	Pharmaceutical and industrial processes	NIR	87
<i>Variance weights (SELECT)</i>	Agriculture	NIR	88
interval, backward and forward-PLS (<i>i-PLS, bi-PLS and fi-PLS</i>)	Green chemistry	NIR	89
Uninformative Variable Elimination-PLS (UVE-PLS)	Food science	Vis-NIR	90
Sequential methods: forward selection and backwards elimination	Mineral science and viticulture	Vis-IR image and chemical analysis	91
Successive projection algorithms (<i>SPA</i>)	Environmental	UV-Vis	92
Genetic Algorithms (<i>GA</i>)-PLS	Remote sensing	UV-Vis-IR and Vis-IR image	93
Self-Organizing maps (<i>SOMs</i>)	Biomedical	NMR	94
Selectivity Ratio	Biomedical	MS	95
Decision trees	Biomedical	DNA microarray	96
Support vector machine (<i>SVM</i>) based-criteria	Biomedical	DNA microarray	97
Wavelets	Polymer industry	Process and quality variables	98

Abbreviations: NIR: Near Infrared; MS: Mass-spectrometry; UV-Vis: Ultraviolet-Visible; IR: Infrared; DNA: Deoxyribonucleic acid; NMR: Nuclear Magnetic Resonance

2.4.4. Data Fusion

The data fusion approach aims to combine data from different sources with the objective to improve the performance of any analysis that it is better than an analysis that uses individual sources alone. A clear example of data fusion is when humans integrate data from their multiple senses (vision, smell and taste) to determine if a certain food product is suitable for consumption. The data fusion process was first developed for military purposes and then further extended to other areas such as robotics, medical diagnosis, environmental monitoring, remote sensing and food processing [99].

Data fusion is increasingly popular because it permits the simultaneous analysis of multiple types of data in order to obtain a global and coherent “perception” of the analytical problem under study. There are different ways of carrying out data fusion. The simplest is to directly fuse the original data, taking into account that it must be correctly balanced (all variables on the same scale) before it is combined. Another way is to fuse a reduced dataset that has been obtained using variable selection and reduction methods. Finally, another strategy is to fuse results obtained from separate multivariate models.

In recent times, several attempts have been made to implement data fusion strategies in various food manufacturing and quality control processes. Table 2.8 gives various examples of some of the spectroscopic techniques most used in data fusion when they are combined with each other or with other devices.

Table 2.8. Examples of data fusion containing spectroscopic techniques in the food ambit.

Techniques	Sample	Study	Reference
NIR/MIR	Extra-Virgin olive oil	Authentication	100
NIR/FT-IR	Cocoa powder	Quantification	101
NMR/IRMS	Mozzarella	Authentication	102
	Cow milk	Authentication	103
NIR/MIR/Fluorescence	Cheese	Authentication	104
HS-MS/FT-IR/UV-Vis	Beer	Characterization	105
HS-MS/Visible/NIR	Wine	Quality attributes	106
AIF/Visible-NIR	Apples	Harvest date and fruit quality	107
e-Tongue/FTIR	Apples	Authentication and quantification	108
e-Tongue/FTIR	Tomatoes	Taste research	109
e-Nose/NIR	Yogurt	Fermentation	110
e-Nose/Visible	Peaches	Quality control	111
Gas sensor/FTIR/UV	White grap must	Authentication	112
Fluorescence images	Apples	Fecal contamination	113

Abbreviations: NIR: Near Infrared; MIR: Mid Infrared; FTIR: Fourier-Transform Infrared;
 NMR: Nuclear Magnetic Resonance; IRMS: Isotope Ratio Mass Spectrometry;
 HS-MS: Head-Space Mass Spectrometry; UV-Vis: Ultraviolet-Visible;
 AIF: Acoustic Impulse Resonance

2.4.5. Multivariate Standardization

A practical limitation that might occur when a multivariate model is already established and validated is that certain changes respect to the initial conditions in which the model was built can diminish its ability to predict future samples. Such changes can arise from:

- Measurements made on different instruments
- Measurements made on the same instrument that are affected by instrumental misalignment, wear, drift, etc.

- Changes in experimental conditions.
- ...

One possible way of dealing with this situation is to construct a new model; however this is not a practical solution because building a robust multivariate model is costly and time consuming process. An attractive alternative is to apply chemometric methods for correcting the analytical signal to adjust it to the initial conditions in which the model was established. The strategies that have been developed to achieve this are called calibration transfer or standardization methods, and are aimed at making an instrument response (for example, the spectra) conform to a “standard” instrument response.

Various methods for standardizing data have been discussed in literature and most of them have been discussed in literature [114, 115]. On the basis of our experience in previous studies, we have decided to use Piecewise Direct Standardization (PDS) in this thesis. This is based on establishing a multivariate relationship between samples measured under different conditions as follows:

$$\mathbf{s}_{1i} = \mathbf{S}_{2[i-j \dots i+j]} \mathbf{b} + \mathbf{b}_0 \quad \text{Eq. (2.6)}$$

where \mathbf{s}_{1i} corresponds to the samples measured under the first condition at the i th variable, \mathbf{S}_2 correspond to the samples measured under the second condition at an interval of variables $[i-j \dots i+j]$, and \mathbf{b} and \mathbf{b}_0 are the coefficients and the additive background correction, respectively, that can be calculated with a multivariate regression method. The coefficients obtained for all variables can be grouped in a diagonal matrix \mathbf{F} according to Figure 2.9:

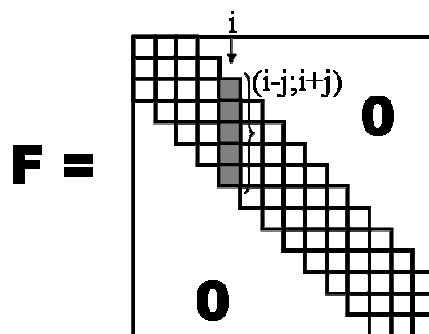


Figure 2.9. Structure of the PDS transformation matrix.

In this matrix, as an example a local model is depicted in bold relating a window of variables around the i th variable (in this case $j=2$) from the second condition to the single i th variable from the initial condition.

Thus the standardization for a sample $(\mathbf{s}_{\text{new}})^{\text{std}}$ is made according to:

$$(\mathbf{s}_{\text{new}})^{\text{std}} = \mathbf{s}_{\text{new}}^T \mathbf{F} + \mathbf{b}_0^T \quad \text{Eq. (2.7)}$$

where S_{new} is the sample measured under the new (second) condition.

Standardization strategies have traditionally been used with a calibration approach, which means that their application to classification problems has been limited. Thus, it can be said that there is plenty of scope for taking advantage of the potential of such methods for use in a classification context. Table 2.9 shows some examples of standardization methods applied to classification.

Table 2.9. Examples of standardization methods used for classification and discrimination.

Standardization method	Pattern Recognition Technique	Scope	Reference
Digital Filtering	PLDA and BNN	Environmental	116
Modified Slope/Bias Correction Orthogonal Signal Correction (OSC) Model updating	LDA	Food Science	117
Standard Normal Variate and Piecewise Direct Standardization (PDS)	PLS-DA, ANN	Agriculture	118
PDS	PCADD	Genetics	119
Modified Direct and Piecewise Standardization (DS and PDS)	PCA	Metabolomics	120
OSC	(<i>k</i> NN)	Medicine	121

Abbreviations: PLDA (Piecewise Linear Discriminant Analysis);

BNN (Back-propagation Neural Network); LDA (Linear Discriminant Analysis);

PLS-DA (Partial Least Squares-Discriminant Analysis); ANN (Artificial Neural Networks);

PCADD (Principal Component Analysis-data description); PCA (Principal Component Analysis);

k NN (*k* -Nearest Neighbours)

2.5. References

- [1] Colour Index, Third Edition. The Society of Dyers and Colourists with the American Association of Textile Chemists and Colourists. Bradford, England (1971).
- [2] P. Gregory, *Dyes Pigm.* **7** (1986) 45-46.
- [3] M.J. Prival, V.M. Davis, M.D. Peiperl, S.J. Bell, *Mutat. Res.* **206** (1988) 247.
- [4] RASFF (Rapid Alert System for Food and Feed), available at:
http://ec.europa.eu/food/food/rapidalert/index_en.htm.
- [5] M. Tripathi, S.K. Khanna, M. Das, *Food Control* **18** (2007) 211.
- [6] ASTA (American Spice Trade Association), available at:
<http://www.astaspice.org/pubs/sudanwhitepaper.pdf>
- [7] *Official Journal of the European Union* (2003/460/EC) L154/114.
- [8] *Official Journal of the European Union* (2004/92/EC) L27/52.
- [9] *Official Journal of the European Union* (2005/402/EC) L135/34.
- [10] C. Baggiani, L. Anfossi, P. Baravalle, C. Giovannoli, G. Giraudi, C. Barolo, G. Viscardi, *J. Sep. Sci.* **32** (2009) 3292.
- [11] C. Ju, Y. Tang, H. Fan, J. Chen, *Anal. Chim. Acta* **621** (2008) 200.
- [12] F. J. López-Jiménez, S. Rubio, D. Pérez-Bendito, *Food Chem.* **121** (2010) 763.
- [13] M.R.V.S. Murty, N. Sridhara Chary, S. Prabhakar, N. Prasada Raju, M. Vairamani, *Food Chem.* **115** (2009) 1556.
- [14] N. Riaz, R. Ali Khan, A.U. Rehman, Z. Ali, S. Yasmeen, N. Afza, *J. Chem. Soc. Pak.* **31** (2009) 151.
- [15] X. Junhong, Z. Dongyan, Z. Limin, X. Zhixiang, Z. Jie, Q. Dejing, *Chromatographia* **73** (2011) 235.
- [16] T. Zhu, Q. Wang, *Afr. J. Agric. Res.* **6** (2011) 1177.
- [17] C. Li, T. Yang, Y. Zhang, Y.L. Wu, *Chromatographia* **70** (2009) 319.
- [18] L. Anfossi, C. Baggiani, C. Giovannoli, G. Giraudi, *Food Addit. Contam.:Part A* **26** (2009) 800.
- [19] C. Li, Y.L. Wu, J.Z. Shen, *Food Addit. Contam.:Part A* **27** (2010) 1215.
- [20] H. Yan, J. Qiao, H. Wang, G. Yang, K.H. Row, *Analyst* **136** (2011) 2629.
- [21] S. Guffogg, P.A. Brown, S.G., Stangroom, C.A. Sutherland, *Bulletin on methods of analysis and sampling for foodstuffs*, Food Standards Agency (FSA) **52** (2004) 26.

- [22] O. Chailapakul, W. Wonsawat, W. Siangproh, K. Grudpan, Y. Zhao, Z. Zhu, *Food Chem.* **109** (2008) 876.
- [23] Y. Ye, B. Xiang, W. Zhang, E. Shang, *Phys. Lett. A* **359** (2006) 620.
- [24] F. Tateo, M. Bononi, *J. Agric. Food Chem.*, **52** (2004) 655.
- [25] R. Rebane, I. Leito, S. Yurchenko, K. Herodes, *J. Chromatogr. A* **1217** (2010) 2747.
- [26] X.C. Chang, X. Z. Hu, Y.Q. Li, Y. J. Shang, Y. Z. Liu, G. Feng, J.P. Wang, *Food Control* **22** (2011) 1770.
- [27] J. Xu, Y. Zhang, J. Yi, M. Meng, Y. Wan, C. Feng, S. Wang, X. Lu, R. Xi, *Analyst* **135** (2010) 2566.
- [28] C.V. Di Anibal, M. Odena, I. Ruisánchez, M.P. Callao, *Talanta* **79** (2009) 887.
- [29] V.S Surwase, K.S. Laddha, R.V. Kale, S.I. Hashmi, S.M. Lokhande, *Electron. J. Environ. Agric. Food Chem.* **10** (2011) 2173.
- [30] L. Maggi, A.M. Sánchez, M. Carmona, C.D. Kanakis, E. Anastasaki, P.A. Tarantilis, M.G. Polissiou, G.L. Alonso, *Food Chem.* **127** (2011) 369.
- [31] D.O. Forcato, M.P. Carmine, G.E. Echeverría, R.P. Pécora, S.C. Kivatinitz, *J. Dairy Sci.* **88** (2005) 478.
- [32] V. Fernández-Ruiz, J.S. Torrecilla, M. Cámara, M.C. Sánchez Mata, C. Shoemaker, *Talanta* **83** (2010) 9.
- [33] X. Zhang, B. Kong, Y.L. Xiong, *Meat Sci.* **77** (2007) 593.
- [34] I. Hussain, L. Khan, G.A. Marwat, N. Ahmed, M. Saleem, *J. Chem. Soc. Pak.* **30** (2008) 406.
- [35] V. León-Ruiz, S. Vera, A.V. González-Porto, M.P. San Andrés, *J. Food Sci.* **76** (2011) C356.
- [36] B. Liu, J. Du, J. Zeng, C. Chen, S. Niu, *Eur. Food Res. Technol.* **230** (2009) 325.
- [37] L. Sun, Y. Zhuang, *Food Bioprocess Technol.* (2010) doi: 10.1007/s11947-010-0406-5.
- [38] A.J. Charlton, *The Food Sector and Nuclear Magnetic Resonance Spectroscopy: A 10 year Overview*, available at:
<http://www.fera.defra.gov.uk/foodDrink/foodAnalysis/aCharltonArticleFeb10.pdf>.
- [39] C.V. Di Anibal, I. Ruisánchez, M.P. Callao, *Food. Chem.* **124** (2011) 1139.
- [40] I. Grong Aursand, E. Veliyulin, U. Erikson, *Mod. Magn. Reson.* **3** (2006) 905.
- [41] B.P. Hills, *Annu. Rep. NMR Spectrosc.* **58** (2006) 177.

- [42] M. Spraul, B. Schütz, E. Humpfer, M. Mörtter, H. Schäfer, S. Koswig, P. Rinke, *Magn. Reson. Chem.* **47** (2009) S130.
- [43] P.S. Belton, A.M. Gil, G.A. Webb, D. Rutledge, *Magnetic resonance in food science: Latest Developments*, The Royal Society of Chemistry, Cambridge, UK (2003) p. 261-266.
- [44] E. Jamin, F. Rique Martin, R. Santamaria-Frenandez, M. Lees, *J. Agric. Food Chem.* **53** (2005) 5130.
- [45] S. Tiziani, S.J. Schwartz, Y. Vodovotz, *J. Agric. Food Chem.* **54** (2006) 6094.
- [46] S.B. Engelsen, P.S. Belton, H.J. Jakobsen, *Magnetic Resonance in Food Science: The Multivariate Challenge*, The Royal Society of Chemistry, Cambridge, UK (2005), p. 72-79.
- [47] K. Van lent, B. Vanlerberghe, P. Van Oostveldt, O. Thas, P. Van der Meeren, *Int. Dairy J.* **18** (2008) 12.
- [48] F. Bertocchi, M. Paci, *J. Agric. Food Chem.* **56** (2008) 9317–9327.
- [49] Y. Ozaki, Raman Spectroscopy. In *Special Methods in Food Analysis: Instrumentation and Applications*, Marcel Dekker Inc. New York (1999) p. 427-462.
- [50] X.F. Zhang, M.Q. Zou, X.H. Qi, F. Liu, X.H. Zhu, B.H. Zhao, *J. Raman Spectrosc.* **41** (2010) 1655.
- [51] S. Gallier, K.C. Gordon, R. Jiménez-Flores, D.W. Everett, *Int. Dairy J.* **21** (2011) 402.
- [52] B.S. Luo, M. Lin, *J. Rapid Meth. Aut. Mic.* **16** (2008) 238.
- [53] A.M. Herrero, *Crit. Rev. Food Sci. Nutr.* **48** (2008) 512.
- [54] M-Q. Zou, X-F. Zhang, X-H. Qi, H-L. Ma, Y. Dong, C-W. Liu, X. Guo, H. Wang, *J. Agric. Food Chem.* **57** (2009) 6001.
- [55] R. Kizil, J. Irudayaraj, *J. Sci. Food Agric.* **87** (2007) 1244.
- [56] J.R. Beattie, S.E.J. Bell, C. Borggaard, B.W. Moss, *Meat Sci.* **80** (2008) 1205.
- [57] S. Thawornchinsombut, J.W. Park, G. Meng, E.C.Y. Li-Chan, *J. Agric. Food Chem.* **54** (2006) 2178.
- [58] A. Skoczowski, M. Troc, A. Baran, M. Baranska, *J. Therm. Anal Calorim.* **104** (2011) 187.
- [59] B. Muika, B. Lendl, A. Molina-Díaz, M. J. Ayora-Cañada, *Chem. Phys. Lipids* **134** (2005) 173.
- [60] S.W. Ellepolaa, S-M. Choib, D.L. Phillipsc, C-Y. Ma, *J. Cereal Sci.* **43** (2006) 85.

- [61] M. Otto, *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, Wiley-VCH Weinheim, Germany (1999) p. 1-2.
- [62] B. Zhu, L. Jiang, F. Jin, L. Qin, A. Vogel, Y. Tao, *Sens. Instrum Food Qual. Saf.* **1** (2007) 123.
- [63] P. Oliveri, M.A. Baldo, S. Daniele, M. Forina, *Anal. Bioanal. Chem.* **395** (2009) 1135.
- [64] Patrycja Ciosek, T. Sobanski, E. Augustyniak, W. Wróblewski, *Meas. Sci. Technol.* **17** (2006) 6.
- [65] N.H. Beltrán, M.A. Duarte-Mermoud, M.A. Bustos, S.A. Salah, E.A. Loyola, A.I. Peña-Neira, J.W. Jalocha, *J. Food Eng.* **75** (2006) 1.
- [66] D. Unay, B. Gosselin, *J. Food Eng.* **78** (2007) 597.
- [67] J.J. Lopez, M. Cobos, E. Aguilera, *Neural Comput. Applic.* **20** (2011) 975.
- [68] M.V. Baroni, C. Arrua, M.L. Nores, P. Fayé, M.P. Díaz, G. A. Chiabrand, D. A. Wunderlin, *Food Chem.* **114** (2009) 727–733.
- [69] S. Ražić, A. Onjia, *Am. J. Enol. Vitic.* **61** (2010) 506.
- [70] R.M. Alonso-Salces, C. Herrero, A. Barranco, D.M. López-Márquez, L.A. Berrueta, B. Gallo, F. Vicente, *Food Chem.* **97** (2006) 438.
- [71] M.L. Busetto, V.M. Moretti, J.M. Moreno-Rojas, F. Caprino, I. Giani, R. Malandra, F. Bellagamba, C. Guillou, *J. Agric. Food Chem.* **56** (2008) 2742.
- [72] O.G. Meza-Márquez, T. Gallardo-Velázquez, G. Osorio-Revilla, *Meat Sci.* **86** (2010) 511.
- [73] H.M. Al-Qadiri, M. Lin, M.A. Al-Holy, A.G. Cavinato, B.A. Rasco, *J. Dairy Sci.* **91**(2008) 950.
- [74] J. Weeranantanaphan, G. Downey, *J. Inst. Brew* **116** (2010) 56.
- [75] I. Stanimirova, B. Üstün, T. Cajka, K. Riddelova, J. Hajslova, L.M.C. Buydens, B. Walczak, *Food Chem.* **118** (2010) 171.
- [76] J.R. Lucio-Gutiérrez, J. Coello, S. Maspoch, *Food Res. Int.* **44** (2011) 557.
- [77] P. Sirisomboon, Y. Hashimoto, M. Tanaka, *J. Food Eng.* **93** (2009) 502.
- [78] F. Shen, Y. Ying, B. Li, Y. Zheng, Q. Zhuge, *Food Chem.* **129** (2011) 565.
- [79] J.S. Ribeiro, T.J. Salva, M.M.C. Ferreira, *J. Food Qual.* **33** (2010) 212.
- [80] Lei Zhang and Lei Nie, *Phytochem. Anal.* **21** (2010) 609.
- [81] L. Della Donna, M. Ronci, P. Sacchetta, C. Di Ilio, B. Biolatti, G. Federici, C. Nebbia, A. Urbani, *Biotechnol. J.* **4** (2009) 1596.

- [82] W. Yuan, B. Xiang, L. Yu, J. Xu, *Food Anal. Methods* (2011), doi:10.1007/s12161-011-9198-0.
- [83] G. Gurdeniz, B. Ozen, *Food Chem.* **116** (2009) 519.
- [84] L. Chen, X. Xue, Z. Ye, J. Zhou, F. Chen, J. Zhao, *Food Chem.* **128** (2011) 1110.
- [85] M. R. Almeida, K. de S. Oliveira, R. Stephani, L. Fernando C. de Oliveira, *J. Raman Spectrosc.* **42** (2011) 1548.
- [86] Y. Saeys, I. Inza, P. Larrañaga, *Bioinformatics* **23** (2009) 2507.
- [87] W. Wu, D.L. Massart, *Chemom. Intell. Lab. Syst.* **35** (1996) 127.
- [88] M. Forina, S. Lanteri, M. Casale, M.C.Cerrato Oliveros, *Chemom. Intell. Lab. Syst.* **87** (2007) 252.
- [89] R.M. Balabina, S.V. Smirnovb, *Anal. Chim. Acta* **692** (2011) 63.
- [90] D. Wu, X. Chen, X. Zhu, X.Guan, G. Wu, *Anal. Methods* **3** (2011) 1790.
- [91] K.Z. Mao, *IEEE T. Syst. Man Cy. B* **34** (2004) 629.
- [92] M.S. Di Nezio, M.F. Pistonesi, W.D. Frago, M.J.C. Pontes, H.C. Goicoechea, M.C.U. Araujo, B.S. Fernández-Band, *Microchem. J.* **85** (2007) 194.
- [93] L. Li, Y-B. Cheng, S. Ustin, X-T. Hu, D. Riaño, *Adv. Space Res.* **41** (2008) 1755.
- [94] G.R. Lloyd, K. Wongravee, C.J.L. Silwood, M. Grootveld, R.G. Brereton, *Chemom. Intell. Lab. Syst.* **98** (2009) 149.
- [95] T. Rajalahti, R. Arneberg, F. S. Berven, K.-M. Myhr, R.J. Ulvik, O.M. Kvalheim, *Chemom. Intell. Lab. Syst.* **95** (2009) 35.
- [96] R. Díaz-Uriarte, S. Alvarez de Andrés, *BMC Bioinformatics* **7** (2006) 1.
- [97] A. Rakotomamonjy, *J. Mach. Learn. Res.* **3** (2003) 1357.
- [98] Y-H. Chu, D. Kim, C. Han, E-S. Yoon, *Ind. Eng. Chem. Res.* **46** (2007) 7188.
- [99] P.M. Ramos, *Raman and X-ray Fluorescence Spectroscopy Data Fusion for Identification of Pigments in Works of Art*, Doctoral Thesis, University Rovira and Virgili, Tarragona (2006).
- [100] M. Casale, N. Sinelli, P. Oliveri, V. Di Egidio, S. Lanteri, *Talanta* **80** (2010) 1832.
- [101] A. Veselá, A. S. Barros, A. Synytsya, I. Delgadillo, J. Copíková, M.A. Coimbra, *Anal. Chim. Acta* **601** (2007) 77.
- [102] M.A. Brescia, M. Monfreda, A. Buccolieri, C. Carrino, *Food Chem.* **89** (2005) 139.
- [103] J.P. Renou, C. Deponge, P. Gachon, J.C. Bonnefoy, J.B. Coulon, J.P. Garel, R. Vérité, P. Ritz, *Food Chem.* **85** (2004) 63.

- [104] R. Karoui, E. Dufour, L. Pillonel, E. Schaller, D. Picque, T. Cattenoz, J.O. Bosset, *Int. Dairy J.* **15** (2005) 287.
- [105] L. Vera, L. Aceña, J. Guasch, R. Boqué, M. Mestres, O. Busto, *Talanta* (2011) doi:10.1016/j.talanta.2011.09.05.
- [106] D. Cozzolino, H.E. Smyth, K.A. Lattey, W. Cynkar, L. Janik, R.G. Damberg, I. Leigh Francis, M. Gishen, *Anal. Chim. Acta* **563** (2006) 319.
- [107] M. Zude, B. Herold, J.M. Roger, V. Bellon-Maurel, S. Landahl, *J. Food Eng.* **77** (2006) 254.
- [108] A. Rudnitskaya, D. Kirsanov, A. Legin, K. Beullens, J. Lammertyn, B.M. Nicolaï, J. Irudayaraj, *Sens. Actuators B* **116** (2006) 23.
- [109] K. Beullens, D. Kirsanov, J. Irudayaraj, A. Rudnitskaya, A. Legin, B. M. Nicolaï, J. Lammertyn, *Sens. Actuators B* **116** (2006) 107.
- [110] M. Navrátil, C. Cimander, C.F. Mandenius, *J. Agric. Food Chem.* **52** (2004) 415.
- [111] C. Di Natale, M. Zude-Sasse, A. Macagnano, R. Paolesse, B. Herold, A. D'Amico, *Anal. Chim. Acta* **459** (2002) 107.
- [112] S. Roussel, V. Bellon-Maurel, J.M. Roger, P. Grenier, *Chemom. Intell. Lab. Syst.* **65** (2003) 209.
- [113] M.S. Kim, A.M. Lefcourt, Y.R. Chen, Y. Tao, *J. Food Eng.* **71** (2005) 85.
- [114] T. Naes, T. Isaksson, T. Fearn, T. Davies, *Multivariate Calibration and Classification*, NIR Publications, Chichester UK (2002) p. 207-219.
- [115] R.N. Feudale, N. A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferré, *Chemom. Intell. Lab. Syst.* **64** (2002) 181.
- [116] F.Koehler, G.W. Small, *Anal. Chem.* **72** (2000) 1690.
- [117] A. Myles, Zimmerman T.A., Brown S., *Appl. Spectrosc.* **60** (2006) 1198.
- [118] S.A. Roussel, C.L. Hardy, C.R. Hurburgh Jr., G.R. Rippke, *Appl. Spectrosc.* **55** (2001) 1425.
- [119] Y. Xu, R.G. Brereton, *Anal. Bioanal. Chem.* **388** (2007) 655.
- [120] T.M. Alam, T.K. Alam, *J. Chemometr.* **24** (2010) 261.
- [121] M. Padilla, A. Perera, I. Montoliu, A. Chaudry, K. Persaud, S. Marco, *Chemom. Intell. Lab. Syst.* **100** (2010) 28.

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

C H A P T E R 3

Experimental Part and Results

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

3.1. Spectroscopic techniques and multivariate analysis for determining Sudan I to IV dyes in culinary spices.

This first part of Chapter 3 contains three papers with one main objective: to explore the potential of using three spectroscopic techniques as screening tools in conjunction with multivariate analysis for determining banned dyes in food matrices.

The first paper (Talanta 79 (2009) 887-892) studies the application of UV-Visible spectroscopy and evaluates the potential of the following three multivariate classification techniques: *K*-Nearest Neighbours (*KNN*), Soft Independent Modelling of Class Analogy (*SIMCA*) and Partial Least Squares-Discriminant Analysis (*PLS-DA*). The performance of each classification technique has been evaluated in terms of classification errors obtained, that is, in terms of false positives and false negatives. The paper discusses the implications of such errors when dealing with a food adulteration problem.

The second paper (Food Chemistry 124 (2011) 1139-1145) studies the use of ^1H -Nuclear Magnetic Resonance (^1H -NMR) as a screening technique. *PLS-DA* was chosen as the classification technique on the basis of previously obtained results. Whenever dealing with high dimensional data as is the case with NMR, a proper variable selection method is required. Consequently, this paper applies a simple method called *Xdiff* to select the most relevant variables. The misclassifications obtained by *PLS-DA* are discussed in terms of samples assigned to more than one class, samples wrongly assigned to a class and samples not assigned to any class.

The third paper (Submitted for publication) investigates the use of three Raman modalities: FT-Raman, normal Raman and SERS. It finds that SERS is the

most appropriate modality for the adulteration problem under study. The background commonly found when working with Raman spectra can mask the Raman signal, so the spectra need to be processed correctly. In this paper, the background interference and the spectral noise are handled by means of two multivariate strategies: Savitzky-Golay smoothing with polynomial baseline correction and wavelet transform. An exploratory analysis (PCA) is applied to raw data and to data processed with the two chemometric treatments in order to determine the discrimination between samples adulterated with Sudan I dye and samples which are free from the adulterant. The paper also makes some suggestions to further improve and extend this methodology.

3.2.1. PAPER

Determining the adulteration of spices with Sudan I-II-III-IV dyes
by UV-Visible spectroscopy and multivariate classification
techniques

Carolina V. Di Anibal, Marta Odena, Itziar Ruisánchez, M. Pilar Callao

Talanta 79 (2009) 887–892

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

Determining the adulteration of spices with Sudan I-II-III-IV dyes by UV-Visible spectroscopy and multivariate classification techniques

Carolina V. Di Anibal¹, Marta Odena², Itziar Ruisánchez¹; M. Pilar Callao¹

¹ Chemometrics, Qualimetrics and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n Tarragona, Spain

² Public Health Laboratory (Catalan Government's Department of Health in Tarragona),
María Cristina 54, 43002 Tarragona, Spain

Abstract

We propose a very simple and fast method for detecting Sudan dyes (I, II, III and IV) in commercial spices, based on characterizing samples through their UV-Visible spectra and using multivariate classification techniques to establish classification rules. We applied three classification techniques: *K*-Nearest Neighbour (KNN), Soft Independent Modelling of Class Analogy (SIMCA) and Partial Least Squares-Discriminant Analysis (PLS-DA). A total of 27 commercial spice samples (turmeric, curry, hot paprika and mild paprika) were analysed by chromatography (HPLC-DAD) to check that they were free of Sudan dyes. These samples were then spiked with Sudan dyes (I, II, III and IV) up to a concentration of 5mg.L⁻¹. Our final data set consisted of 135 samples distributed in five classes: samples without Sudan dyes, samples spiked with Sudan I, samples spiked with Sudan II, samples spiked with Sudan III and samples spiked with Sudan IV.

Classification results were good and satisfactory using the classification techniques mentioned above: 99.3%, 96.3% and 90.4% of correct classification with PLS-DA, KNN and SIMCA respectively. It should be pointed out that with SIMCA, there are no real classification errors as no samples were assigned to the wrong class: they were just not assigned to any of the pre-defined classes.

Keywords: Sudan dyes; KNN; PLS-DA; SIMCA; Spices; Multivariate analysis.

1. Introduction

Sudan I (1-[(2,4-dimethylphenyl) azo]-2-naphthalenol), Sudan II (1-(phenylazo)-2-naphthol), Sudan III (1-(4-phenylazophenylazo)-2-naphthol) and Sudan IV (o-tolyazo-o-tolyazo-betanaphthol) are an azo-family of synthetic dyes that are widely used for colouring agents such as waxes, floor and shoe polishes. They are categorized as class 3 carcinogens by the International Agency for Research on Cancer (IARC) [1]. As a result, Sudan dyes are illegal as additives in foodstuffs destined for human consumption according to both the FSA (Food Standards Agency) [2] and the European Union. Unfortunately, in some countries, these dyes are still being used as additives in some foodstuffs to improve the colour for commercial benefits. Reports have indicated that high amounts of Sudan dyes, at least 1 g.L^{-1} , are required to have an impact on visual colour [3].

Several methods have been proposed to detect the presence of some of these synthetic dyes in foodstuffs: for example, high performance liquid chromatography-diode array detection HPLC-DAD [4,5], gel permeation chromatography-mass spectrometry [6], pressurized capillary electrochromatography (CEC) with amperometric detection [7], HPLC with electrochemical detection [8], liquid chromatography-mass spectrometry, [9,10,11],

tandem mass spectroscopy and isotope dilution [12], multi-wall carbon nanotube-based electrochemical sensing [13], and the ELISA method [14]. Also, second order multivariate techniques that use a data matrix for each sample have been applied to determine Sudan I in chilly powder [15]. In most of these methods, time-consuming pre-separations are often needed, analysis is not so fast and the instrumentation is not so affordable.

Multivariate techniques have been successfully applied to the analysis of foodstuffs. Some recent papers on this issue are: the multivariate prototype approach for authenticating red wines [16], the classification of milks according to their origin [17], the authentication of salmon and salmon-based products [18], characterisation and discrimination among butters according to their fat content [19], the authentication and classification of olive oil [20,21,22], the classification of vinegar [23] and the classification of apple fruits and ciders [24,25].

In this study, we propose a very simple and fast method for detecting Sudan dyes in commercial spices, based on characterizing samples through their UV-Visible spectra and using multivariate classification techniques to establish classification rules. Five classes are considered: class 1: samples without Sudan dyes, class 2: samples spiked with Sudan I, class 3: samples spiked with Sudan II, class 4: samples spiked with Sudan III and class 5: samples spiked with Sudan IV. Therefore, an unknown sample (the possible Sudan dye content of which is unknown) is assigned to one of the five classes considered in this study.

First of all, we applied an exploratory analysis based on the well known Principal Component Analysis (PCA) technique [26,27] in order to detect natural sample grouping with no previous information. Then, we applied classification techniques, both hard (discriminating) and soft (modelling) techniques. In the first, the hyperspace is divided in as many regions as the number of existing classes so, if

a sample falls in the region of space corresponding to a particular class; it is classified as belonging to only this class. In the last, frontiers are built between each class and the rest of the space [28], modelling each class separately. The decision rule for a given class is a class box that envelopes the position of the class and an object can be assigned in more than one class or not assigned to a class [29]. We applied *K*-Nearest Neighbours (*KNN*) [30,31], Soft Independent Modelling of Class Analogy (*SIMCA*) [32,33] that is a soft classification technique and Partial Least Squares Discriminant Analysis (*PLS-DA*) that are hard classification techniques. *PLS-DA* is a variant of *PLS* [34] in which instead of predicting a quantitative parameter a qualitative assignation (class) is done [35,36,37].

We are dealing with a contamination problem, so it is important to consider the type of classification error because the associated cost and practical implications are different. From a practical point of view, then, stating that a sample does not contain Sudan dye when it does is not the same as stating that a sample contains Sudan dye when it does not. In this scenario, we will present the results on the basis of the following null hypothesis, H_0 : "Sample does not contain Sudan dye". Therefore, as it is shown in Table 1, type I and type II errors will be studied when each classification technique is applied. Type I errors might have economic implications as samples will be removed from the market with no real need. The consequences of type II errors are worse as they involve a health risk: consumers will buy samples contaminated with Sudan dye. Type I and II errors are related with sensitivity and specificity [38].

To the best of our knowledge, the one-step determination of the four Sudan dyes on the basis of their UV-Visible spectra and multivariate classification has yet to be reported.

Table 1. Type of errors made according to the decision taken satisfying the accomplishment of a null hypothesis.

	H0 TRUE	H0 FALSE
Statistical decision: reject H_0	<i>Type I error</i>	<i>Correct decision</i>
Statistical decision: do not reject H_0	<i>Correct decision</i>	<i>Type II error</i>

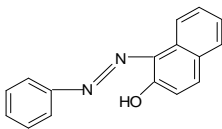
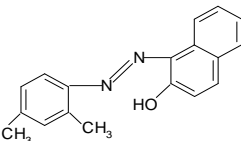
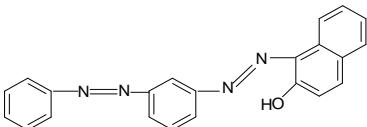
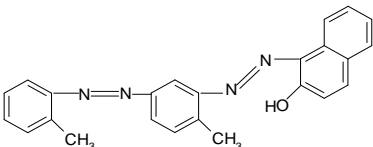
2. Experimental

2.1. Reagents and samples

Table 2 shows the chemical structure of the four Sudan dyes. It can be seen that all four Sudan dyes have some common structure but that Sudan III and IV have an additional azo and benzene group. Also, Sudan II and IV have two methyl groups more than Sudan I and III. The Sudan I standard was purchased from ACROS (Geel, Belgium) and the other Sudan dyes were purchased from SIGMA (St. Louis, MO, USA). A total of 27 spices from different commercial trade were purchased from markets. Acetonitrile and chloroform (for HPLC and UV-Visible analysis) and acetic acid (for HPLC) were all of HPLC grade.

Samples contaminated with Sudan dye were obtained from the non-contaminated samples. To prepare the working solutions that were to be used to spike the commercial spices, appropriate amounts of Sudan I, II, III and IV were diluted in acetonitrile. Sudan III and Sudan IV were initially dissolved in a small fraction of chloroform and then diluted with acetonitrile.

Table 2. Chemical structure of Sudan dyes.

Dye	Structure
Sudan I	
Sudan II	
Sudan III	
Sudan IV	

The samples to be analysed, either by HPLC or UV-Vis, followed an extraction process: one gram of sample was weighed and 50 mL of acetonitrile was added. The sample was extracted with magnetic stirring. Each extract was obtained by filtering twice, first with glass microfibre filters and then with syringe filters. In order to obtain the UV-Visible spectra, a volume of this extract (350 μL for mild and hot paprika, 300 μL for curry and 70 μL for turmeric) was spiked with the appropriate amount of one Sudan dye in such a way that the final concentration was 5 mgL^{-1} . Table 3 contains the list of the commercial samples analyzed, the corresponding spiked samples and the numbers assigned to them all.

Table 3. Description of the samples studied: commercial name and assigned number.

Spice	Commercial Trade	Original and spiked spices				
		Original	Sudan I	Sudan II	Sudan III	Sudan IV
Turmeric	Hacendado	1	28	55	82	109
	Jugosan	2	29	56	83	110
	Corbella	3	30	57	84	111
	non-branded	4	31	58	85	112
	non-branded	5	32	59	86	113
Curry	Eroski	6	33	60	87	114
	Carmencita	7	34	61	88	115
	Bonpreu	8	35	62	89	116
	Caprabo	9	36	63	90	117
	Hacendado	10	37	64	91	118
	Dia	11	38	65	92	119
Mild paprika	Carmencita	12	39	66	93	120
	Bonpreu	13	40	67	94	121
	Caprabo	14	41	68	95	122
	Hacendado	15	42	69	96	123
	Carrefour	16	43	70	97	124
	"De la Vera" Carrefour	17	44	71	98	125
	Dani	18	45	72	99	126
	Gourmet	19	46	73	100	127
Hot paprika	Eroski	20	47	74	101	128
	Carmencita	21	48	75	102	129
	Bonpreu	22	49	76	103	130
	Caprabo	23	50	77	104	131
	Hacendado	24	51	78	105	132
	Dani	25	52	79	106	133
	Gourmet	26	53	80	107	134
	"De la Vera" El Reu	27	54	81	108	135

2.2. Apparatus and software

The HPLC is a Varian Star (Varian Inc., USA) model equipped with the following modules: a Prostar 240 pump; a Prostar 410 automatic injector and a Prostar 335 diode array detector (DAD). The relevant parameters were: column, Agilent Zorbax ODS (Agilent Technologies, USA) 250mm x 4.6mm (5 μ m particle size); mobile phase, acetonitrile/acid acetic 16% (70:30, v/v) in gradient mode; flow rate 1.0mL min⁻¹; temperature 40°C. Photometric detection was

performed at 478 nm for Sudan I and 510 nm for Sudan II, III and IV. The software used was the STAR 6.41 Chromatography Workstation.

The UV-Visible spectrophotometer (Agilent 8453, United States) was equipped with a diode array detector (DAD). UV-visible spectrum scanning was carried out in the wavelength range of 260-600 nm (each nm) which represents 340 variables.

The data measured was processed with Matlab 6.5 software (Version 6.5, The Math Works Inc., Natick, USA) and PLS Toolbox 3.5 (Eigenvector Research Incorporated). Data was pre-processed with mean centering before each chemometric treatment.

4. Results and Discussion

Figures 1.a and 1.b show, respectively, the spectra of a standard solution of the four Sudan dyes at 5 mgL^{-1} and one spectrum for each of the four spices studied (turmeric, curry, hot paprika and mild paprika), randomly chosen. It can be seen that the spectra of the four Sudan dyes have different shapes although there is considerable overlap. Sudan I and Sudan II (as well as Sudan III and Sudan IV) have similar spectra according to their chemical structure. Mild and hot paprika present similar spectra shape at the same wavelength interval, as do turmeric and curry but with a slight blue-shift of the absorption peak.

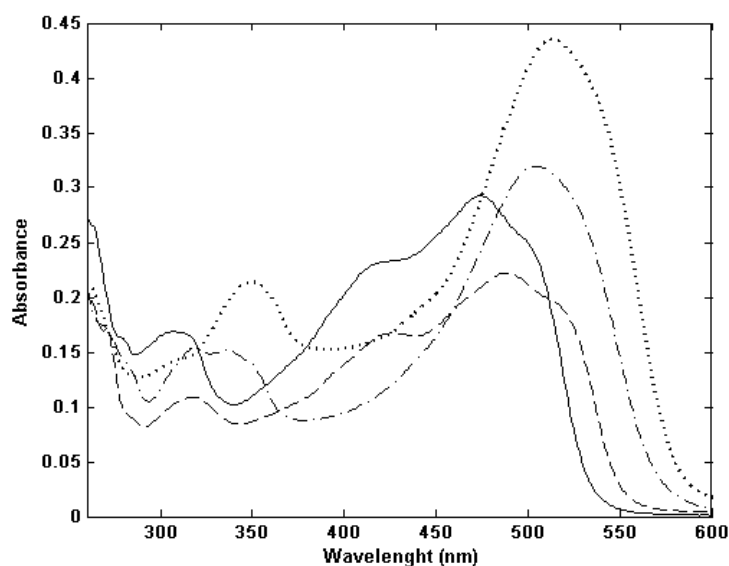


Figure 1.a. UV-Visible spectra of Sudan I, II, III and IV at 5 ppm ($\text{mg}\cdot\text{L}^{-1}$). Sudan I (solid line), Sudan II (dashed line), Sudan III (dotted-dashed line) and Sudan IV (dotted line).

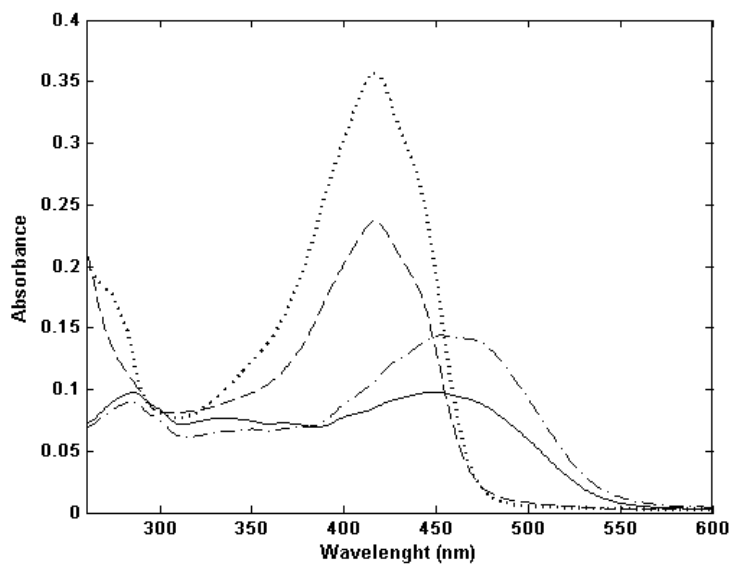


Figure 1.b. UV-Visible spectra of turmeric sample (dotted line), curry sample (dashed line) hot paprika sample (solid line) and mild paprika sample (dashed-dotted line).

Figures 2.a and 2.b show the HPLC chromatograms of a mixture of the four Sudan standards and a random paprika sample, respectively. It should be pointed out that all paprika samples have similar behaviour. Sudan I, II and IV are not present in the sample analyzed as there is no chromatographic peak at the expected retention time. There is some uncertainty, however, about the presence of Sudan III in the paprika samples as the sample chromatogram has a peak at its expected retention time. When turmeric and curry samples are analysed, there is only one chromatographic peak but its retention time is lower than that of the peaks associated with any of the four Sudan dyes. In this case, these samples are free of any kind of Sudan.

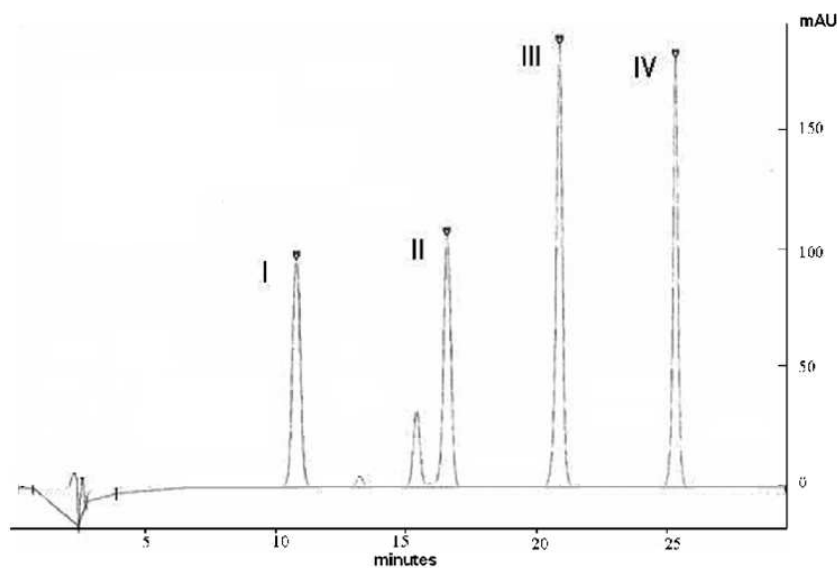


Figure 2.a. Chromatogram of Sudan standard dyes at 5 mg.L^{-1} .

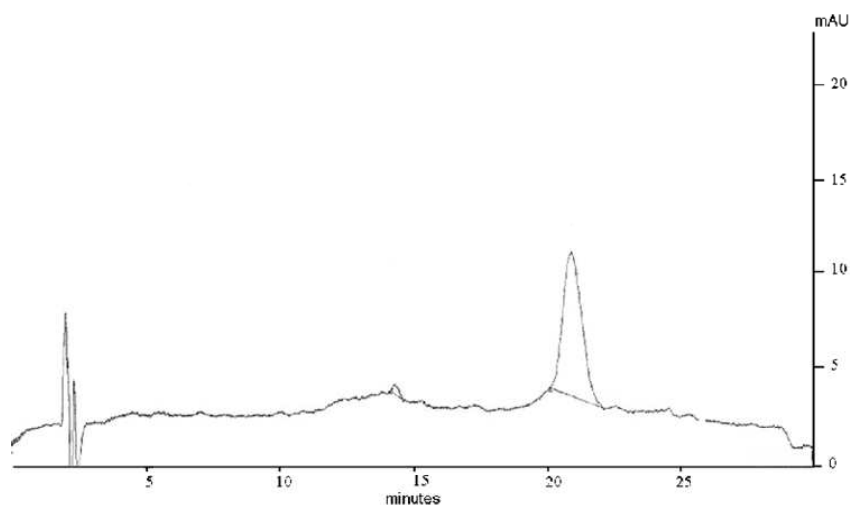


Figure 2.b. Chromatogram of mild paprika (sample n° 15).

First of all, an exploratory analysis of the 135 UV-Visible spectra corresponding to samples indicated in Table 3 was made by principal component transformation. The PCA scores plot is shown in Figure 3, with 86.13% of the total variance explained by the first two principal components. The plot shows that samples that do not contain Sudan are well separated from those that do, with the exception of samples n° 17 and n° 27 which overlap the samples that contain Sudan II. These samples are a special variety of mild and hot paprika called “De la Vera”. It should be stressed that none of the samples without Sudan are placed in the group of samples that contain Sudan III, which indicates that, as expected, the commercial samples analysed do not contain Sudan III.

The samples spiked with the Sudan dyes tended to be grouped according to the type of dye they contain, but the different classes are not clearly separated. Also, in each of the five classes considered, there is a grouping corresponding to the type of spice. So, in the case of samples without Sudan, curries (n° 6 to 11) and turmeric (n° 1 to 5) are spread along PC2, and have similar PC1 score values, whereas the group corresponding to mild (n° 12 to 19) and hot paprika (n° 20 to 27) are more

compact and have similar PC2 score values. Likewise, samples spiked with the four Sudan dyes have a similar behaviour that is translated linearly to the different classes.

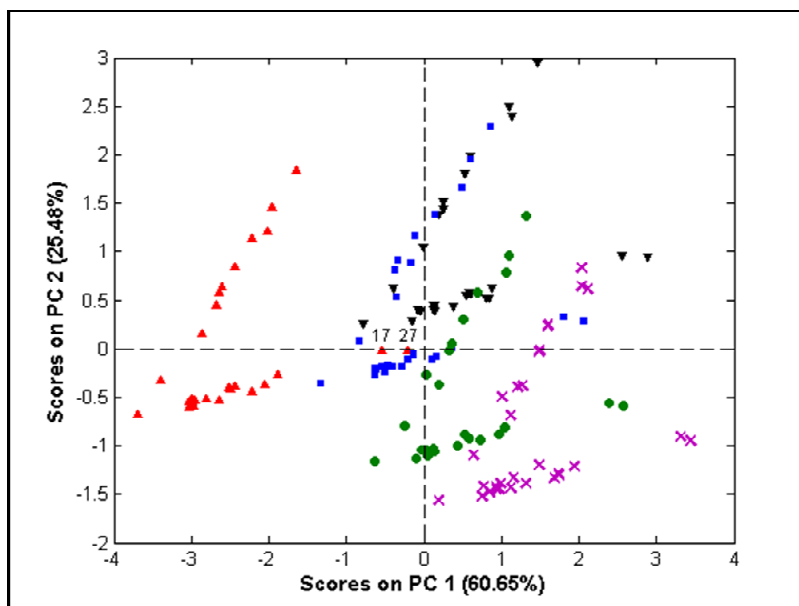


Figure 3. Scores plot of commercial unspiked samples (triangles up) and samples spiked at 5 mg L^{-1} with Sudan I (triangles down), Sudan II (squares), Sudan III (circles) and Sudan IV (crosses).

KNN, SIMCA and PLS-DA were used to make the classifications to the five categories. The models built during the study were validated using the leave-one-out cross validation approach [39].

The KNN classification rule is based on the similarity between neighbors calculated from the Euclidean distance. Therefore, an unknown sample is classified to the class to which belongs the majority of k neighbours closer to it. The classification percentage for each class is the parameter used to determine how many neighbours (k) need to be considered. The results were best when five neighbours

were considered. The only exception was class 1 (samples without Sudan dyes) in which the classification results were best with three neighbours. So $k=5$ was always considered for all the classes in the classification step.

Similarly, the number of principal components required to build the SIMCA model in each of the five pre-defined classes, is determined by the maximum prediction ability. Table 4 shows the details of the SIMCA model for each class and the number of PC's retained for each class are indicated in bold.

Table 4. SIMCA model details: number of principal components (PC's) and percentage of correct classifications and predictions. The values corresponding to the n° of PC's chosen for each model are shown in bold.

	N° of principal components	Classification ability	Prediction ability
Class 1: unspiked spices	1	88.9	85.2
	2	96.3	81.5
	3	96.3	92.6
	4	96.3	85.2
Class 2: spices spiked with Sudan I	1	37.0	18.5
	2	92.6	85.2
	3	92.6	85.2
Class 3: spices spiked with Sudan II	1	18.5	7.4
	2	92.6	85.2
	3	96.3	85.2
	4	96.3	88.9
	5	96.3	88.9
Class 4: spices spiked with Sudan III	1	33.3	18.5
	2	92.6	88.9
	3	92.6	85.2
	4	96.3	96.3
	5	96.3	92.6
Class 5: spices spiked with Sudan IV	1	29.6	29.6
	2	92.6	88.9
	3	92.6	88.9

The SIMCA approach enables the T^2 Hotelling and Q statistics to be studied. T^2 gives a measure of the fit of each sample to the obtained model, while Q is a measure of the residuals, and is therefore related to the sample information not included in the model. The T^2 and Q limits allow establishing the sample relation with the model [40]. In the five classes studied, some samples are out of the limits of T^2 : samples n° 27 for class 1, n° 44 and n° 54 for class 2, and n° 125 and n° 135 for class 5 and some samples are out the limits of Q : n° 1 for class 1, n° 33 for class 2, n° 60 and 77 for class 3, n° 87 for class 4 and n° 114 for class 5. Sample n° 62 is at the limit of the class to which it belongs (class 3). Finally, almost all samples belonging to class 3 (samples which contain Sudan II), are close to the limits of class 2 (samples which contain Sudan I).

According to the best prediction ability (see Table 5), the PLS-DA model was performed using six latent variables. The PLS-DA model consists of building a conventional PLS model, but with class indicator variables as Y matrix. PLS-DA is carried out using an exclusive binary coding scheme with one bit per class, so to discriminate between five classes, a response encoded $\{1,0,0,0,0\}$ means that the sample belongs to class 1, to discriminate between five classes, a response and a response encoded $\{0,1,0,0,0\}$ means that the sample belongs to class 2, and so on. During the modelling process, the PLS-DA method is trained to compute the five “membership values”, one for each class. The class assignment result is expressed in terms of a value that is normally distributed around 0 when the prediction is that the sample is not in a class and around 1 when the prediction is that it is in a class. For this reason it is possible to obtain prediction values below 0 (negative values) and also above 1.

Table 5. PLS-DA model details: number of latent variables (LVs), percentage of cumulative variance for the X and Y block and prediction ability. The values corresponding to the number of LVs chosen are shown in bold.

N° of LVs	%Cumulative variance		Prediction ability				
	X-block	Y-block	Class 1	Class 2	Class 3	Class 4	Class 5
1	60.10	16.15	92.6	11.1	81.5	44.4	85.1
2	84.01	30.02	92.6	81.5	70.4	81.5	92.6
3	96.41	36.97	92.6	96.3	77.8	81.5	100
4	99.14	41.75	100	96.3	77.8	71.4	100
5	99.58	57.25	100	96.3	85.2	81.5	100
6	99.79	69.94	100	100	96.3	100	100
7	99.79	78.59	100	100	96.3	100	100

PLS-DA uses the distribution of calibration-sample predictions to determine a threshold value based on the Bayes Theorem, which will best split those classes with the least probability of false classifications for future predictions. It is assumed that the predicted values for each class are approximately normally distributed. From this distribution the probability of a sample to belong to each class is calculated and is then assigned to the class that has the highest probability value.

Figure 4 shows the PLS-DA assigned values for each sample and the threshold value for each class (presented as horizontal lines). For the sake of clarity, only samples with an assignment value higher than a threshold value are depicted. According to those values, there are two samples assigned to more than one class. One is sample n° 85 (non branded turmeric spiked with Sudan III), which belongs to class 4 and it is assigned to its own class and to class 3. In that case, the probability values are 88% and 52%, respectively, so this sample is considered as belonging to class 4. The other is sample n° 114 (Curry Eroski spiked with Sudan IV), which belongs to class 5 and it is assigned to its own class and to class 4. In that case, the probability values are 100% and 98%, respectively, so, assignment to only one of the predefined classes is not possible. Finally, sample 60 (Curry Eroski

spiked with Sudan II,)) which belongs to class 3 has an assignment value above the threshold value for class 4 and the probability value is 89%, so it is wrongly assigned as belonging to class 4.

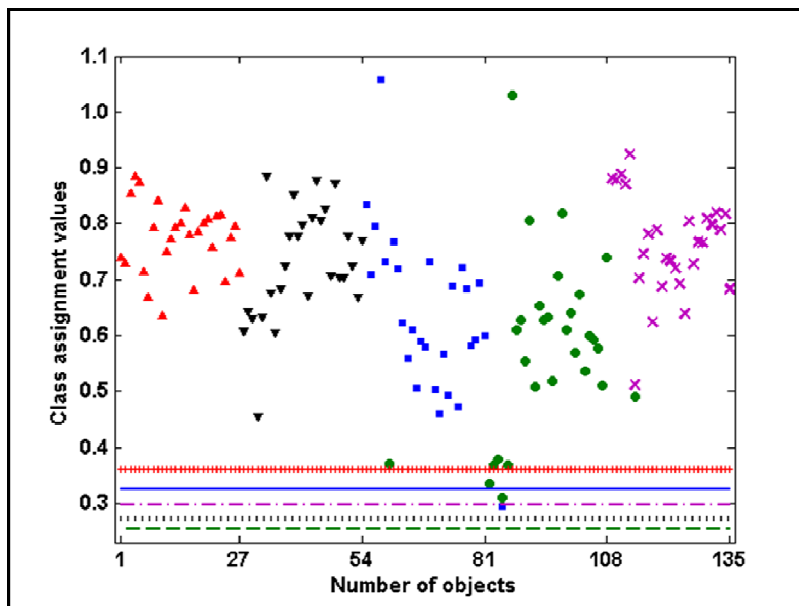


Figure 4. Assigned class values VS objects number according to table 3. Class 1, samples without Sudan dyes (triangles up); class 2, samples spiked with Sudan I (triangles down); class 3, samples spiked with Sudan II (squares); class 4, samples spiked with Sudan III (circles); class 5, samples with Sudan IV (crosses). Threshold values: class 1 (vertical dashed line); class 2 (dotted line); class 3 (solid line); class 4 (dashed line) and class 5 (dotted-dashed line).

Finally, Table 6 shows the class assignments for each sample when the three classification techniques are used. The wrongly classified samples are shown in bold. The sample numbers are the same as the ones in Table 3.

From the results, several overall conclusions can be drawn: a) No samples containing Sudan dye (samples in classes 2, 3, 4 and 5) are assigned to class 1 (samples that do not contain Sudan dye). This is of great importance from the point of view of the type of error, which will be discussed later; b) the SIMCA approach does not assign any samples to the wrong class, although it does not assign 13 samples to any class at all. In our opinion this is a clear advantage over the other classification techniques (*KNN*, for instance, will always assign a sample to one of the predefined classes; c) the classification results, in terms of classification percentage, were best with the PLS-DA approach (only sample n° 60 was wrongly assigned).

Another general trend is that all samples spiked with Sudan dye from the non-contaminated Curry sample called "Eroski" (samples n° 33, n° 60, n° 87 and n° 114) were not assigned to any class when the SIMCA approach was used. Sample n° 60 (which contains Sudan II) was wrongly assigned when the PLS-DA approach was used. With the mild and hot paprika samples called "De la Vera", the original and some spiked samples are not assigned or wrongly assigned. This might be due to a special behaviour or sample composition. For the other wrongly assigned samples, no general trends are observed.

From the point of view of the null hypothesis, H_0 ("The analyzed sample is not contaminated. It does not contain Sudan") and based on Table 1, the following situations have to be discussed. The null hypothesis was rejected when it was true for two samples when using the *KNN* approach (n° 17 and n° 27), two samples when using SIMCA (n° 1 and n° 27) and no samples when using PLS-DA

approach. So with the first two techniques, type I error is 7.4%. From a practical point of view, the normal procedure with samples that are detected as contaminated is to submit them to a confirmatory analysis (in our case HPLC-MS). This means additional time and cost. In our particular case, the key error is type II error, which means that the sample is said to be not contaminated, that it does not contain Sudan, when it does. Our results show that we have 0% of type II error, which is important because no confirmatory method will ever be applied.

Table 6. Class assignment for each sample with *KNN*, *SIMCA* and *PLS-DA*.

Wrongly assigned samples are indicated in bold.

	Assignment to					Not assigned to any class
	Class 1	Class 2	Class 3	Class 4	Class 5	
PLS-DA	1-27	28-54	55-59	60 , 82-108	109-135	
<i>KNN</i>	1-16	28-30	17, 27, 31	82-108	109-135	
	18-26	32-54	55-64			
		65, 71	66-70			
			72-81			
<i>SIMCA</i>	2-26	28-30, 32, 34	55-59, 61	82-86	109-113	1, 27, 31, 31, 33
		35-43	63-76	88-108	115-124	44, 54, 60, 62,
		45-53	78-81		126-134	77, 87, 114
					125, 135	

5. Conclusions

The use of UV-Visible spectroscopy with multivariate classification techniques is a simple and inexpensive methodology for determining the possible contamination of foodstuffs, such as culinary spices, with Sudan dye I to IV.

Considering the null hypothesis approach, the consequences of wrongly assigning a contaminated sample as non-contaminated (type II error) might involve a health risk. It should be pointed out that the results are extremely positive as no

wrong assignments are made whenever one of the three classification techniques is applied.

A comparison between the three classification techniques shows that the results were best with PLS-DA, as only one sample was wrongly assigned. What is more, this wrong assignment involves no health or economic risk as it is a contaminated sample assigned to a different group of contaminated samples. The *KNN* and *SIMCA* approaches assign some non-contaminated samples as contaminated. Finally, three contaminated samples are wrongly assigned by *KNN* because of a mistaken Sudan assignment. The *SIMCA* approach does not assign some samples to any class. In these cases a confirmatory method would be required.

When the paprika samples without Sudan dyes (class 1) were subject to chromatographic analysis, it could not be stated whether Sudan III was present or not (class 4). With the proposed methodology, however, it is clear that Sudan III is not present because no samples without Sudan are within or close to the samples that contain Sudan III.

Acknowledgments

The authors would like to thank the Spanish Ministry of Education, Culture and Sports (Project CTQ2007-61474/BQU) for economic support, and the Management Agency for University and Investigation Support of the Catalan Government for providing Carolina Di Anibal's doctoral fellowship.

6. References

- [1] IARC (1975) Monographs on the evaluation of the carcinogenic risk of chemical to man: some aromatic azo compounds (Vol. 8). Lyon, France, International Agency for Research on Cancer, pp. 224-231.
- [2] FSA (Food Standards Agency):
<http://www.food.gov.uk/foodindustry/guidancenotes/foodguid/sudanguidance>
- [3] America Spice Trade Association (ASTA):
<http://www.astaspice.org/pubs/sudanwhitepaper.pdf>
- [4] M. Mazzetti, R. Fascioli, I. Mazzoncini, G. Spinelli, I. Morelli, A. Bertoli, *Food Addit. Contam. Part A* **21** (2004) 935.
- [5] M. Ma, X. Luo, B. Chen, S. Su, S. Yao, *J. Chromatogr. A* **1103** (2006) 170.
- [6] H. Sun, F. Wang, L. A., *J. Chromatogr. A* **1164** (2007) 120.
- [7] L. Shaofeng, Z. Xue, L. Xucong, W. Xiaoping, F. Fengfu, X. Zenghong, *Electrophoresis* **28** (2007) 1696.
- [8] O. Chailapakul, W. Wonsawat, W. Siangproh, K. Grudpan, Y. Zhao, Z. Zhu, *Food Chem.* **109** (2008) 876.
- [9] H. Limin, S. Yijuan, F. Binghu, S. Xiangguang, Z. Zhenling, L. Yahong, *Anal. Chim. Acta* **594** (2007) 139.
- [10] O. Pardo, V. Yusà, N. León, A. Pastor, *Talanta* **78** (2009) 178.
- [11] M.R.V.S. Murty, N. Sridhara Chary, S. Prabhakar, N. Prasada Raju, M. Vairamani, *Food Chem.* **115** (2009) 1556.
- [12] L. Di Donna, L. Maiuolo, F. Mazzotti, D. De Luca, G. Sindona, *Anal. Chem.* **76** (2004) 5104.
- [13] T. Gan, K. Li, K. Wu, *Sens. Actuators B* **132** (2008) 134.
- [14] Y. Wang, D. Wei, H. Yang, Y. Yang, W. Xing, Y. Li, A. Deng, *Talanta* **77** (2009) 1783.
- [15] Y. Jintao, L. Lifu, L. Yingwu, D. Changai, H. Bo, *Anal. Chim. Acta* **607** (2008) 160.
- [16] S. Preys, E. Vigneau, G. Mazerolles, V. Cheynier, D. Bertrand, *Chemom. Intell. Lab. Syst.* **87** (2007) 200.
- [17] P. Ciosek, W. Wróblewski, *Talanta* **76** (2008) 548.
- [18] S. Masoum, C. Malabat, M. Jalali-Heravi, C. Guillou, S. Rezzi, D.N. Rutledge, *Anal. Bioanal. Chem.* **387** (2007) 1499.
- [19] B. Jaillais, V. Morrin, G. Downey, *Chemom. Intell. Lab. Syst.* **86** (2007) 179.

- [20] M. Casale, C. Casolino, G. Ferrari, M. Forina, *J. Near Infrared Spectrosc.* **16** (2008) 39.
- [21] F. Fernandes Gambarra-Neto, G. Marino, M.C. Ugulino Araújo, R. Kawakami Harrop Galvão, M.J. Coelho Pontes, E.P. de Medeiros, R. Sousa Lima, *Talanta* **77** (2009) 1660.
- [22] M. Casale, C. Armanino, C. Casolino, M. Forina, *Anal. Chim. Acta* **589** (2007) 89.
- [23] M. Casale, C. Armanino, C. Casolino, C. Oliveros, M. Forina, *Food Sci. Technol. Res.* **12** (2006) 223.
- [24] R.M. Alonso-Salces, C. Herrero, A. Barranco, L. Berrueta, B. Gallo, F. Vicente, *Food Chem.* **93** (2005) 113.
- [25] R.M. Alonso-Salces, S. Guyot, C. Herrero, L. Berruela, J. Drilleau, B. Gallo, F. Vicente, *Food Chem.* **91**(2005) 91.
- [26] D.L. Massart, , B.G.M. Vandeginste, L.M.C Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics, Part A*, Elsevier, Amsterdam, The Netherlands (1997).
- [27] R.G. Brerenton., *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, John Wiley & Sons, Chichester, UK (2003).
- [28] D. González-Arjona, A.G. González, *Anal. Chim. Acta* **363** (1998) 89.
- [29] M.P. Derde, D.L. Massart, *Mikrochim. Acta* **II** (1986) 139.
- [30] R. Duda, P. Hart, D. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, New York, 2000 p. 20-84.
- [31] W. Wu, D. Massart, *Anal. Chim. Acta* **349** (1997) 253.
- [32] D.L. Massart, , B.G.M. Vandeginste, L.M.C Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics, Part B*, Elsevier, Amsterdam, The Netherlands (1998).
- [33] S. Wold, M. Sjöström, *Chemometrics: Theory and Application*, In: B.R. Kowalski (Ed.), ACS Symposium Series No. 52, American Chemical Society, Washington D.C., USA (1977).
- [34] P. Geladi, B. Kowalski, *Anal. Chim. Acta* **185** (1986) 1.
- [35] F. Liu, Y. He, L. Wang, *Anal. Chim. Acta* **615** (2008) 10.
- [36] M. Barker, W. Rayens, *J. Chemom.* **17** (2003) 166.
- [37] K. Hovde Liland, U. Geir Nidal, *J. Chemom* **23** (2009) 7.
- [38] M.C. Ortiz, L. Sarabia, R. García-Rey, M.D. Luque de Castro, *Anal. Chim. Acta* **558** (2006) 125.

[39] Stone M., *J. R. Sta. Soc. B* **36** (1974) 111.

[40] A. Rius, M.P. Callao, F.X. Rius, *Analyst* **122** (1997) 737.

3.2.2. PAPER

High-Resolution ^1H -Nuclear Magnetic Resonance spectrometry
combined with chemometric treatment to identify adulteration of
culinary spices with Sudan dyes

Carolina V. Di Anibal, Itziar Ruisánchez, M. Pilar Callao

Food Chemistry 124 (2011) 1139-1145

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

High-Resolution ^1H -Nuclear Magnetic Resonance spectrometry combined with chemometric treatment to identify adulteration of culinary spices with Sudan dyes

Carolina V. Di Anibal, Itziar Ruisánchez, M. Pilar Callao

Chemometrics, Qualimetrics and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n Tarragona, Spain.

Abstract

An efficient method for detecting the adulteration of commercial culinary spices with Sudan I, II, III and IV dyes is proposed using a combination of High resolution ^1H -Nuclear Magnetic Resonance and chemometric treatment. The variables were reduced and selected on the basis of the difference between the NMR spectra from the non contaminated commercial spices and the spices spiked with one of the four Sudan dyes. Partial Least Squares-Discriminant Analysis (PLS-DA) was applied to the most important NMR variables selected. The commercial spices studied were curry, turmeric, and mild and hot paprika, distributed in five classes: non contaminated spices and spices spiked independently with one Sudan dye. The prediction probabilities provided by PLS-DA were satisfactory for all the classes. Only one spiked sample was misclassified in another contaminated class and it should be stressed that no spices from any of the contaminated classes is assigned to the non contaminated one. This is very important from the point of view of consumer health, since a suspicious sample which might contain Sudan dyes will be correctly recognized as adulterated.

Keywords: Nuclear Magnetic Resonance; Sudan dyes; Spices adulteration; PLS-DA; Chemometrics; Variable selection.

1. Introduction

Commercial spices are increasingly included in a range of meals that can be prepared by the consumer to enhance flavor and aroma and to create variety. Sudan I, II, III and IV are fat soluble azo-dyes that are added to food, such as culinary spices, to intensify and maintain its natural red appearance over time. These dyes are classified by the International Agency for Research on Cancer [1] as class 3 carcinogens, but they have recently been found in some foodstuffs in some European countries. In response to this situation, the European Commission [2] requires products to have documentation confirming the absence of Sudan dyes, so there is a demand for reliable and accurate analytical methods for the fast determination of such compounds in foodstuffs.

A search of the scientific bibliography indicates that a large number of techniques have been developed for analysing these dyes in foodstuffs. One of the most commonly used is High Performance Liquid Chromatography (HPLC) with UV detection [3,4], but other detection systems have also been used, such as mass spectrometry [5,6], fluorescence [7] and chemiluminescence [8]. Other methods include reversed-phase liquid chromatography-electrospray-tandem mass spectrometry [9], capillary electrophoresis [10], solid phase spectrophotometry [11], a combination of UV-Visible spectroscopy and multivariate techniques [12], plasmon resonance light scattering [13] and electrochemical reduction at a glassy carbon electrode [14].

In the present study we propose a method that combines High Resolution ^1H Nuclear Magnetic Resonance (^1H -NMR) and chemometric data analysis to detect whether commercial culinary spices have been adulterated with Sudan dyes (Sudan I, II, III and IV). High-resolution ^1H -NMR spectrometry is well suited to the analysis of complex samples because it is reproducible, it requires minimal sample, sample preparation is not time consuming and it is noninvasive or minimally invasive. In recent years, rapid developments in instrument design have resulted in significant improvements in both resolution and sensitivity [15,16]. Also, chemometric methods are increasingly being applied to the processing and extraction of information from large datasets such as NMR data [17,18].

High Resolution Nuclear Magnetic Resonance spectrometry has been combined with multivariate analysis and applied to a variety of processes with foodstuffs: for example, the authentication of olive oil [19], honey [20] and fruit juice [21], the characterization of feedstock [22], the detection of chemical contamination in soft drinks [23], the adulteration of orange drinks [24] and olive oil [25] and the quality assessment of green tea [26].

When NMR measurements are used with multivariate analysis, there may be a problem of high dimensionality. An NMR spectrum has several thousand measurement points, only a few of which contain relevant information, so if the performance of multivariate techniques is to be improved, a subset of variables should be selected. In the present work, first the NMR spectra were collected and then a univariate approach was used to select the variables.

Among the large number of multivariate classification techniques, there is a well known supervised pattern recognition technique known as Partial Least Squares-Discriminant Analysis (PLS-DA). This technique works very well for data with small sample sizes and a large number of variables. PLS-DA uses a few latent

variables rather than a lot of measured variables, which means that it has several benefits. It takes into account the correlation among the variables, filters noise and leads to a good predictive performance, especially when it is combined with variable selection methods. PLS-DA in conjunction with NMR analysis has been successfully applied to determine different kinds of adulteration and contamination of foodstuffs [19,27,28].

To the best of our knowledge, no reports have previously been published on the one-step detection of Sudan dyes in foodstuffs by NMR spectrometry.

2. Materials and Methods

2.1. *Samples and chemicals*

Commercial spices (27 samples) of different trademarks were obtained from common markets, distributed as follows: turmeric (5), curry (6), mild paprika (8) and hot paprika (8). Deuterated chloroform for NMR analysis (99.8 at % D) was provided by SDS (Carlo Erba Reagents SDS S.A., Spain). Sudan I standard was provided by ACROS (Geel, Belgium) and Sudan II, III and IV were provided by SIGMA (St. Louis, MO, USA).

2.2. *Preparation of the commercial and spiked samples*

Prior to the NMR analysis, High Performance Liquid Chromatography with a Diode Array Detector (HPLC-DAD) was used to confirm that all the commercial spices were free of any Sudan dye [12].

Stock solutions of all Sudan dyes were prepared in deuterated chloroform and stored in the refrigerator. The commercial samples to be analysed were prepared by weighing 0.1 g of each spice, dissolving it in 5 mL of deuterated chloroform and filtering using syringe nylon filters (45 μm). From this solution, 700 μL was taken and placed in 2 mL flasks.

Spiked samples were prepared by adding the stock Sudan (I to IV) solutions to each commercial sample to obtain a final concentration of 50 mgL^{-1} . This concentration is within the range of adulterated species [29,30].

So, the final samples number is 135 distributed in five classes of 27 samples each: samples 1 to 27 were the commercial ones (class 1), 28 to 54 the ones spiked with Sudan I (class 2), 55 to 81 the samples spiked with Sudan II (class 3), 82 to 108 the ones spiked with Sudan III (class 4) and, finally, 109 to 135 the ones spiked with Sudan IV (class 5).

2.3. NMR analysis

The ^1H -NMR experiments were performed on a Varian NMR System 400. ^1H resonances were detected at 400.13 MHz using a 4 μs pulse (45°), an acquisition time of 2.2 s (32,768 complex points) and a spectral width of 7217 Hz (18 ppm). Sixteen scans were recorded per sample to produce data with optimized sensitivity. Free induction decay (FID) data were Fourier transformed (FT), with prior phase and baseline correction, and converted into ASCII files for further analysis. The data were processed using Mestre-C software (Version 2.3a). Chemical shifts are expressed in δ scale (ppm) and referenced to the residual signal of chloroform (7.26 ppm) [31]. The spectra were measured from 0.5 to 8.8 ppm, so the final number of variables was 8471.

2.4. Data analysis

2.4.1. NMR variable selection

To select the NMR variables, we calculate the x_{diff} values in accordance with Eq. 1 [23].

$$x_{diff,ij} = \frac{|x_{ij} - \bar{x}_i|}{\sigma_i} \quad (1)$$

where x_{ij} is the NMR intensity at frequency i for sample j , and \bar{x}_i and σ_i are the mean and the standard deviation of the intensity, respectively, at each frequency i for the 27 commercial samples not contaminated with any Sudan dye.

The X_{diff} matrix is then obtained from all the samples studied ($j=1$ to 135) and for all frequencies ($i=1$ to 8471). Its magnitude is indicative of the frequency intensities corresponding to the Sudan dyes because a x_{ij} value much greater than \bar{x}_i means that this particular frequency has information that is not present in the non contaminated samples, so it could be an indication of the presence of a Sudan dye. Therefore, x_{diff} values were used to rank the NMR spectra for feature selection. Then, a threshold value was defined from the x_{diff} values calculated for class 1 (natural commercial spices) and only those frequencies (variables) with x_{diff} values higher than the prefixed threshold were selected.

2.4.2. PLS-DA multivariate data analysis

Although it was developed as a regression method, PLS [32] can be used to solve classification problems, in which case it is known as PLS-DA and it encodes

the class membership of the measured samples in the target matrix [33,34]. An important feature of this supervised technique is that it is specifically suited to deal with problems in which the number of variables is large (compared to the number of observations) and collinear, two major challenges encountered when ^1H NMR data are used.

In PLS-DA, a PLS regression model is calculated that relates the independent variables (spectral data) to a binary response encoded as: $\{1,0,0,0,0\}$ means that sample belongs to class 1, $\{0,1,0,0,0\}$ means that belongs to class 2, and so on until class 5. A threshold value based on the Bayesian method [35], is defined between 0 and 1, and calculated on the basis of the values predicted during the classification process, so an object is assigned to a particular class if its prediction is larger than the threshold value for this class. As in the PLS regression model, the optimal number of latent variables (LVs) retained must be chosen before the modelling process and this is done using the root mean square error cross validation (RMSECV), in terms of the fractional misclassification error rate. The classification process is evaluated by the leave-one-out cross-validation approach [36].

2.5. Software

Each NMR spectrum was exported as an ASCII file to Matlab. Statistical (variable selection) and chemometric analyses (data processing and PLS-DA) were made under the Matlab environment (Version 6.5, The Mathworks, Natick, USA) and PLS Toolbox 3.5 (Eigenvector Research Incorporated).

3. Results and Discussion

3.1. NMR spectroscopy analysis

^1H -NMR spectra were recorded of the 27 natural commercial spices and of the corresponding samples spiked with Sudan I, II, III and IV. Figures 1.a to 1.d show the NRM spectra of the four types of spice samples studied and figures 1.e to 1.h show the spectra of the four pure Sudan dyes (I to IV). The four types of spices present quite similar spectrum. The spectrum of the Sudan dyes present most of the signals in the aromatic and methyl regions. It can be seen that the signals of the species appear highly overlapped with the Sudan signals. In a preliminary study, in which all the variables are used, the PLS-DA classification results (in percentages) were very poor. With the exception of class 3, they were always lower than 50% (48% for class 1, 44% for class 2, 59% for class 3, 11% for class 4 and 33% for class 5). This result is somehow expected, as it is well known that the inclusion of too many variables without relevant information regarding the problem ahead is not adequate to carry out a multivariate analysis. These preliminary results show that variables must be selected in order to detect the possible presence of Sudan dyes in foodstuffs destined for human consumption. As a first trial, a visual selection based on the spectral analysis of the analysed samples was carried out and an improvement of the classification results was obtained. Therefore, the use of a mathematical criterion is studied to properly select the variables to build the PLS-DA model that give the best classification results.

3.2. Variable selection

Figure 2 shows the X_{diff} matrix representation for each class in accordance with Equation 1. It can be seen from the non contaminated samples (class 1) that

x_{diff} values minimize the difference between the samples and variables (whatever the kind of spice or the variable considered), all of which are found in a narrow interval. For the rest of the spiked classes, however, the x_{diff} values are high in some of the zones studied and depend on the adulterant they contain. To select the optimal threshold value, several values were checked. Table 1 shows the global percentage classification error obtained from PLS-DA for each threshold value, and the number of variables selected. The classification error is minimum for a threshold value of 5.5 (bold values), which is a significant variable reduction (around nine times lower), so this is the threshold value that was finally chosen.

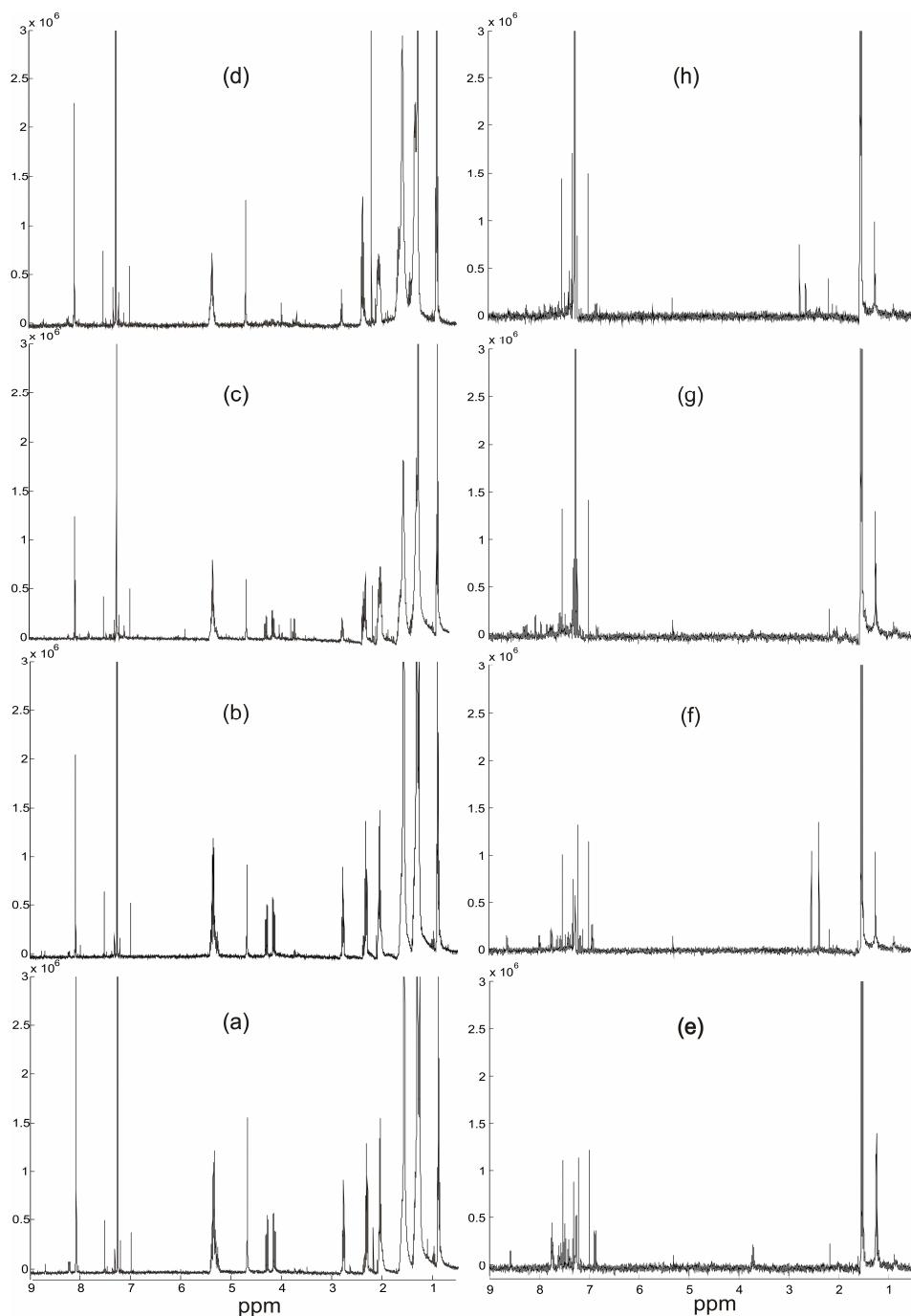


Figure 1. NMR spectra of one the different types of studied spices (left) an the four Sudan dyes (right): hot paprika (a), mild paprika (b), curry (c), turmeric (d), Sudan I (e), Sudan II (f), Sudan III (g) and Sudan IV (h).

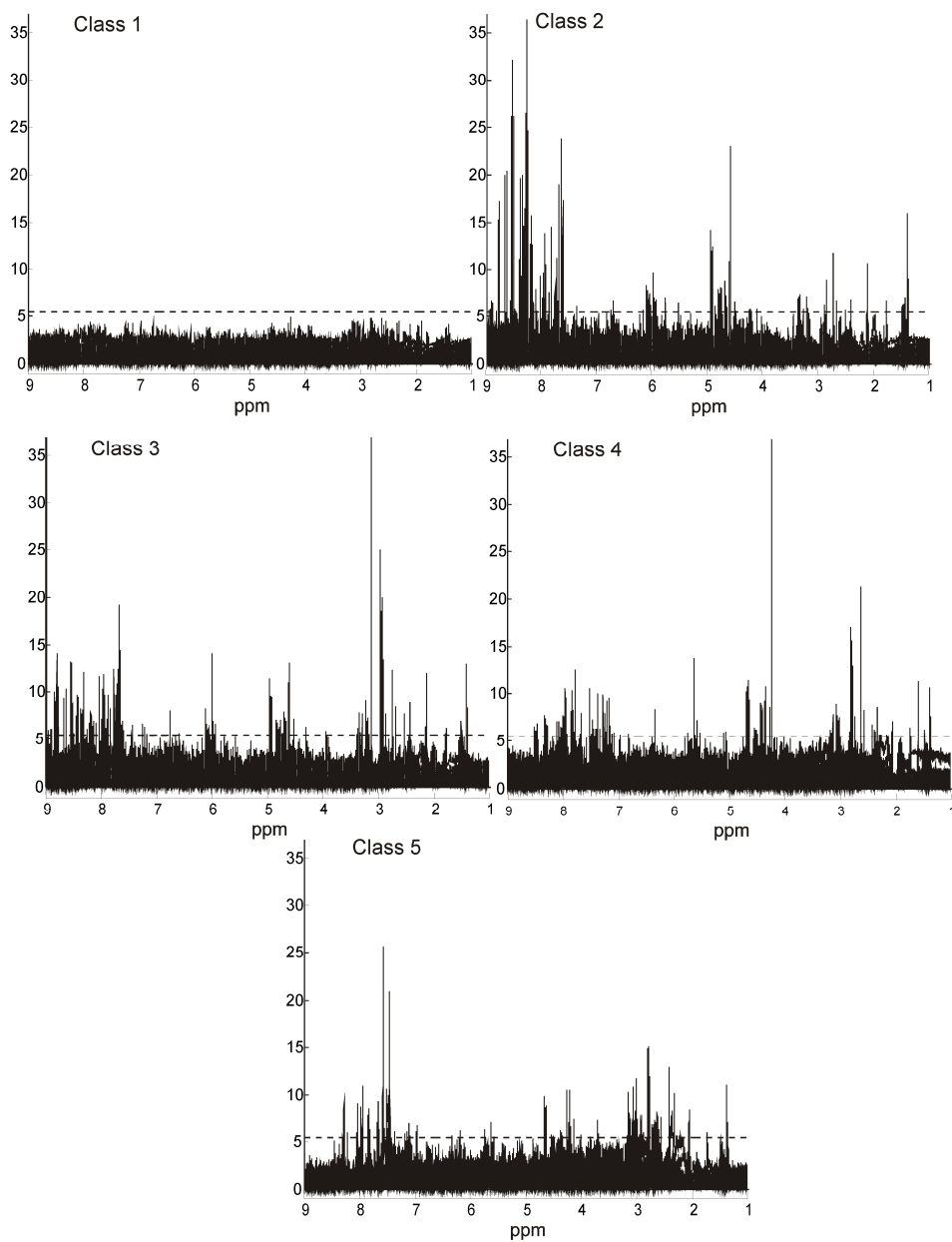


Figure 2. X_{diff} values for class 1 (non contaminated samples), class 2 (samples spiked with Sudan I), class 3 (samples spiked with Sudan II), class 4 (samples spiked with Sudan III) and class 5 (samples spiked with Sudan IV).

Table 1. Number of selected variables and overall PLS-DA classification error (%) for the thresholds tested. Bold values correspond to the threshold value chosen.

Threshold value	N° of selected variables	Global classification error (%)
4.5	1933	5.9
5.0	1389	4.4
5.5	999	3.0
6.0	798	6.7
6.5	644	6.7

Figures 3.a and 3.b show, respectively, the original and the reduced spectrum for a mild paprika spiked with Sudan I. Considerable reduction can be seen in the aliphatic zone between 0.8 and 3 ppm. Also, between 4 and 6 ppm there is a slight reduction, mainly for the multiplet at approximately 5.4 ppm. In the last part of the studied spectrum (the aromatic zone), which is between 7 and 9 ppm, there is no considerable NMR signal reduction because this zone is enriched when Sudan dyes are present in the sample (see Figure 1).

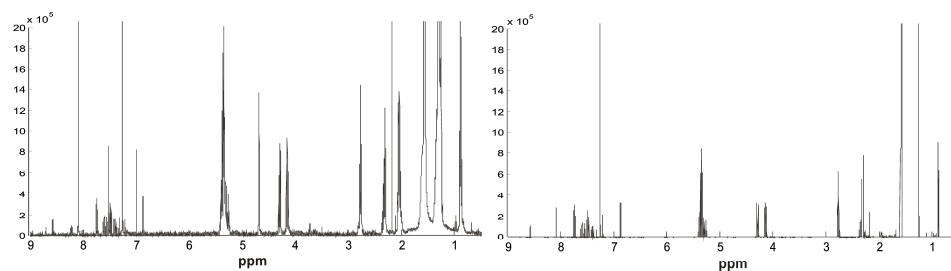


Figure 3. Spectrum with all the original variables (a) and reduced spectrum after variable selection (b) for mild paprika spiked with Sudan I.

3.3. PLS-DA results

The final data set is a 135 x 999 matrix, the rows of which represent the samples (each class has 27 rows) and the columns the standardized NMR intensities (x_{diff}) selected by the procedure outlined above. As has been stated previously, the optimal number of LVs used to build the PLS-DA model is selected on the basis of the RMSECV values. This number is a commitment between the optimal LVs numbers obtained for each individual class as the PLS-DA methodology does not allow choosing the LVs for each class, so the same LVs are used for all classes. Table 2 shows the percentage of recognition and prediction for each class obtained with different number of LVs. From these results, 8 LVs was the number chosen to build the final PLS-DA model. These results can be considered somehow optimistic since the PLS-DA model has been validated by leave-one-out cross-validation analysis instead of building the model with a training and test set, which is the optimal case if the available samples are enough and there is not much variability among the samples, which is not our case.

Table 2. Percentage of recognition (%R) and prediction (%P) for each class obtained with PLS-DA models varying the number of latent variables (LVs). Bold values correspond to the number of LVs finally chosen.

Number of LVs	Class 1		Class 2		Class 3		Class 4		Class 5	
	% R	% P	% R	% P	% R	% P	% R	% P	% R	% P
5	88.9	88.9	92.6	92.6	96.3	96.3	66.7	55.5	70.4	63
6	96.3	96.3	92.6	92.6	96.3	96.3	88.9	88.9	81.5	77.8
7	96.3	96.3	100	100	96.3	96.3	88.9	88.9	85.2	81.5
8	96.3	96.3	100	100	96.3	96.3	92.6	92.6	92.6	92.6
9	96.3	96.3	100	100	96.3	96.3	92.6	92.6	92.6	92.6
10	100	100	100	100	96.3	96.3	96.3	96.3	96.3	96.3

Table 3 shows some information regarding the 23 samples with different classification results. In this table, the probability of belonging to their own class is always shown while the predicted probability of belonging to another class is only shown for those samples with a probability value higher than 2%.

Table 3. Sample information and PLS-DA predicted probability (%) for samples with some kind of exception to the general behaviour of all samples.

Sample number	Spice	Trade mark	True class	Predicted probability (%)				
				Class 1	Class 2	Class 3	Class 4	Class 5
55	turmeric	Hacendado	3		100	1	55	
<u>122</u>	<u>mild paprika</u>	<u>Caprabo</u>	<u>5</u>				<u>9</u>	<u>39</u>
2	turmeric	Jugosan	1	100				80
107	hot paprika	Gourmet	4				100	71
127	mild paprika	Gourmet	5				83	100
97	mild paprika	Carrefour	4				96	95
5	turmeric	non-branded	1	92			2	
27	hot paprika "de la Vera"	El Reu	1	96			11	
33	curry	Eroski	2		100		6	
43	mild paprika	Carrefour	2		100		11	
47	hot paprika	Eroski	2		100			10
75	hot paprika	Carmencita	3			100	36	
81	hot paprika "de la Vera"	El Reu	3			100	15	
96	mild paprika	Hacendado	4				100	16
98	mild paprika "de la Vera"	Carrefour	4				100	22
99	mild paprika	Dani	4				98	25
100	mild paprika	Gourmet	4				100	29
101	hot paprika	Eroski	4				100	9
103	hot paprika	Bonpreu	4				100	7
106	hot paprika	Dani	4				100	55
112	turmeric	non-branded	5				11	80
119	curry	Dia	5				9	100
133	hot paprika	Dani	5				18	97

A detailed look at Table 3 shows that only sample n° 55 is assigned with a very low probability (1%) to its true class (class 3), and with maximum probability to class 2 (100%). There is also some probability that it belongs to class 4 (55%). Only one sample (n° 122, underlined values in table 2) was not assigned to any of

the five predefined classes. It belongs to class 5, but the probability of being assigned to this class is rather low (39%) although higher than the probability of being assigned to the other classes. In this case, in our opinion, the assignation is not possible.

The four other samples marked in italics (Table 3) have the highest probability of belonging to their own class but they also have a high probability of belonging to another class. Samples n° 2, n° 107 and n° 127 have a 100% probability of belonging to its true class but also a high probability (80, 71 and 83%, respectively) of belonging to a different class. Slightly different is the case of sample n° 97, since it has the same probability of belonging to both class 4 (96%, true class) and class 5 (95%), so in that case its assignation to one of the contaminated classes is not clear. In the particular case of sample 2, an uncontaminated spice is properly assigned with 100% to its own class, but it also would be suspected of being contaminated with Sudan IV because there is a high probability (80%) that it belongs to class 5. So to confirm the status of this sample, a confirmatory analysis should be done, such as High Performance Liquid Chromatography (HPLC) [29,37]. As far as the other samples mentioned above are concerned, the possibility of being assigned to another class does not have the same importance as sample n° 2, because they are contaminated samples that may be classified in another contaminated class (classes 4 or 5).

The other samples in Table 3 that have not been mentioned can be clearly assigned to their true class rather than to another class, considering the classifications probabilities in each of them. It can be seen that most of them are between class 4 and 5, spiked with Sudan III and IV, respectively. This may be because the characteristic NMR spectrum of Sudan III is similar to the spectrum of Sudan IV (see Figure 1): they have the same common chemical structure and only differ in two methyl groups [12]. This, however, is not the case for class 2 and class

3, which are spiked with Sudan I and Sudan II. Although they also have a very similar chemical structure and again only differ in two methyl groups, there is no misclassification between both classes.

It should be emphasised that no samples from any of the contaminated classes are assigned to the non contaminated class, so the specificity of class 1 is 100% and there were no false negatives: i.e., spices adulterated with Sudan dyes will not be approved for human consumption.

4. Conclusions

This investigation demonstrates that $^1\text{H-NMR}$ spectrometry coupled with appropriate multivariate statistical analysis is an efficient technique for determining the adulteration of commercial spices with Sudan I, II, III and IV azo dyes, which are prohibited as additives in foodstuffs by European legislation.

This study shows the need to reduce and select variables, since a lot of the thousand of NMR frequencies registered do not have relevant classification information and using all of them worsen the results. The variable selection used based on the calculation of *Xdiff* matrix is a proper method to select variables that allow distinguishing between the five sample groups under study.

The proposed methodology was demonstrated with culinary spices destined for human consumption. PLS-DA results were highly satisfactory since only one sample from a contaminated class is misclassified as another contaminated class. It should be pointed out that the four doubtful classifications are contaminated samples assigned to more than one contaminated class, but never to the non-contaminated class. None of the contaminated spices with one of the four dyes

were assigned to the non-contaminated category. This is very important for consumer health, since a suspicious sample which might contain Sudan dyes will be correctly recognized as adulterated.

Acknowledgements:

The authors would like to thank Marta Odena from the Public Health Laboratory in Tarragona and the Management Agency for University and Investigation Support of the Catalan Government (AGAUR) for providing Carolina Di Anibal a doctoral fellowship and the Spanish Ministry of Education, Culture and Sports (Project CTQ2007-61474/BQU) for economic support.

5. References

- [1] IARC. (1975). Monographs on the evaluation of the carcinogenic risk of chemicals to man: some aromatic azo compounds (Vol. 8). Lyon, France: International Agency for Research on Cancer, pp. 224–231.
- [2] Commission Decision (2005). Commission decision 2005/402/EC of 23 May 2005 on emergency measures regarding chilli, chilli products, curcuma and palm oil. Official Journal of the European Communities, L135, 34–36.
- [3] V. Cornet, Y. Govaert, G. Moens, J. Van Loco, J.M. Degroodt, *J. Agri. Food Chem.* **54** (2006) 639.
- [4] D.G. Hussein, P.A. Biacs, *J. Chromatogr. Sci.* **43** (2005) 461.
- [5] F. Tateo, M. Bononi, *J. Agri. Food Chem.* **52** (2004) 655.
- [6] M. van Bruijnsvoort, S.J.M. Ottink, K.M. Jonker, E. de Boer, *J. Chromatogr. A* **1058** (2004) 137.
- [7] A. Pielesz, I. Baranowska, A. Rybak, A. Wochowicz, *Ecotoxicol. Environ. Saf.* **53** (2002) 42.
- [8] Y. Zhang, Z. Zhang, Y. Sun, *J. Chromatogr. A* **1129**(2006) 34.
- [9] F. Calbiani, M. Careri, L. Elviri, A. Mangia, L. Pistarà, I. Zagnoni, *J. Chromatogr. A* **1042** (2004) 123.
- [10] E. Mejia, Y. Ding, M.F. Mora, C.D. Garcia, *Food Chem.* **102** (2007) 1027.
- [11] F. Capitán, L.F. Capitán-Vallvey, M.D. Fernández, I. de Orbe, R. Avidad, *Anal. Chim. Acta* **331** (1996) 141.
- [12] C.V. Di Anibal, M. Odena, I. Ruisánchez, M.P. Callao, *Talanta*, **79** (2009) 887.
- [13] L.P. Wu, Y.F. Li, C.Z. Huang, Q. Zhang, *Anal. Chem.* **78** (2006) 5570.
- [14] M. Du, X. Han, Z. Zhou, S. Wu, *Food Chem.* **105** (2007) 883.
- [15] E. Danieli, J. Mauler, J. Perlo, B. Blümich, F. Casanova, *J. Magn. Reson.* **198** (2009) 80.
- [16] H.C. Jarrell, *J. Magn. Reson.* **198** (2009) 204.
- [17] L.A. Berrueta, R.M. Alonso-Salces, K. Héberger, *J. Chromatogr. A* **1158** (2007) 196.
- [18] J.C. Lindon, E. Holmes, J.K. Nicholson, *Prog. Nucl. Magn. Reson. Spectrosc.* **39** (2001) 1.
- [19] R.M. Alonso-Salces, K. Héberger, M.V. Holland, J.M. Moreno-Rojas, C. Mariani, G. Bellan, F. Reniero, C. Guillou, *Food Chem.* **118** (2010) 956.

- [20] J.A. Donarski, S.A. Jones, M. Harrison, M. Driffield, A.J. Charlton, *Food Chem.* **118** (2010) 987.
- [21] M. Cuny, E. Vigneau, G. Le Gall, I. Colquhoun, M. Lees, D.N. Rutledge, *Anal. Bioanal. Chem.* **390** (2008) 419.
- [22] G.E. Pereira, J.P. Gaudillere, C. Van Leeuwen, G. Hilbert, O. Lavialle, M. Maucourt, C. Deborde, A. Moing, D. Rolin, *J. Agric. Food Chem.* **53** (2005) 6382.
- [23] A.J. Charlton, P. Robb, J.A. Donarski, J. Godward, *Anal. Chim. Acta*, **618** (2008) 196.
- [24] G. Le Gall, M. Puaud, I.J. Colquhoun, *J. Agric. Food Chem.* **49** (2001) 580.
- [25] D.L. García-González, L. Mannina, M. D'Imperio, A.L. Segre, R. Aparicio, *Eur. Food Res. Technol.* **219** (2004) 545.
- [26] G. Le Gall, I.J. Colquhoun, M. Defernez, *Journal of Agricultural and Food Chemistry* **52** (2004) 692–700.
- [27] E.F. Boffo, L.A. Tavares, M.M.C. Ferreira, A.G. Ferreira, *LWT- Food Sci. Technol.* **42** (2009) 1455.
- [28] L. Tarachiwinl, O. Masako, E. Fukusaki, *J. Agric. Food Chem.* **58** (2008) 5827.
- [29] ASTA (2005). <<http://www.astaspice.org/files/public/SudanWhitePaper.pdf>>.
- [30] K. Mishra, S. Dixit, S.K. Purshottam, R.C. Pandey, M. Das, S.K. Khanna, *Int. J. Food Sci. Technol.* **42** (2007) 1363.
- [31] R.E. Hoffman, *Magn. Reson. Chem.* **44** (2006) 606.
- [32] P. Geladi, B. Kowalski, *Anal. Chim. Acta*, **185** (1986) 1.
- [33] M. Barker, W. Rayens, *J. Chemom.* **17** (2003) 166.
- [34] L. Stahle, S. Wold, *J. Chemom.* **1** (1987) 185.
- [35] R. Duda, P. Hart, D. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, New York, 2000 p. 20-84.
- [36] L. Kryger, *Talanta* **28** (1981) 871-887.
- [37] FSA (Food Standards Agency):
<<http://www.food.gov.uk/foodindustry/guidancenotes/foodguid/sudanguidance>>.

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

3.2.3. PAPER

Surface Enhanced Raman Spectroscopy (SERS) and multivariate analysis as a screening tool for detecting Sudan I dye in culinary spices

Carolina V. Di Anibal, Lluís F. Marsal, M. Pilar Callao, Itziar Ruisánchez

Submitted

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

Surface Enhanced Raman Spectroscopy (SERS) and multivariate analysis as a screening tool for detecting Sudan I dye in culinary spices

Carolina V. Di Anibal¹, Lluís F. Marsal², M. Pilar Callao¹, Itziar Ruisánchez¹

¹Chemometrics, Qualimetrics and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n Tarragona, Spain

²Nanoelectronic and Photonic Systems, Department of Electronic, Electric and Automatic Control Engineering, Rovira i Virgili University, Avda. Països Catalans 26, 43007, Tarragona, Spain

Abstract

Raman spectroscopy combined with multivariate analysis was evaluated as a tool for detecting Sudan I dye in culinary spices. Three Raman modalities were studied: normal Raman, FT-Raman and SERS. The results show that SERS is the most appropriate modality capable of providing a proper Raman signal when a complex matrix is analyzed. To get rid of the spectral noise and background, Savitzky-Golay smoothing with polynomial baseline correction and wavelet transform were applied. Finally, to check whether unadulterated samples can be differentiated from samples adulterated with Sudan I dye, an exploratory analysis such as Principal Component Analysis (PCA) was applied to raw data and data processed with the two mentioned strategies. The results obtained by PCA show that Raman spectra need to be properly treated if useful information is to be obtained and both spectra treatments are appropriate for processing the Raman signal. The proposed methodology shows that SERS combined with appropriate spectra treatment can be

used as a practical screening tool to distinguish samples suspicious to be adulterated with Sudan I dye.

Keywords: SERS, Sudan I, Spices, Food adulteration, Multivariate analysis

1. Introduction

Sudan I dye is a synthetic red dye traditionally used for coloring plastics and textile products. Because of its low cost and wide availability, this dye is also attractive as food colorant, and it has also been used to reinforce the natural color of some culinary spices. Sudan I is considered to be a class 3 carcinogen (IARC, International Agency for Research on Cancer) and genotoxic [1], so it is prohibited as a food additive according to the European framework [2]. Numerous techniques have been used to detect Sudan I dye in food products, most of which involve the use of chromatographic techniques. A detailed review of these analytical techniques is given elsewhere [3].

In recent years, there has been increasing interest in the use of vibrational spectroscopic methods such as Raman spectroscopy to evaluate food safety and quality [4-9]. This technique offers a rapid analysis with a wealth of chemical information about the composition of a sample that can be used for qualitative or quantitative purposes, requires minimal sample preparation and can perform measurements in any state (solid, liquid, gas). Also, it is a versatile technique and has several modalities, each one with a different purpose: Surface Enhanced Raman Spectroscopy, Resonance Raman Spectroscopy, Coherent anti-Stokes Raman Spectroscopy, among others. The most important drawback of Raman arises when the samples produce a fluorescent background that obscures the weak Raman signal. These aspects need to be minimized so that the spectra are consistent.

One of the Raman modalities that has developed considerably over the last few years is Surface-enhanced Raman Scattering (SERS). Its potential can be appreciated by the increasing number of studies in the literature, for example, in the field of food products [10-15]. With SERS, Raman measurements are made by depositing the sample onto metallic colloidal or metallic solid substrates of nanoscale roughness. These metallic nanostructures enhance the Raman signal intensity by several orders of magnitude due to two mechanisms: the chemical enhancement in which charge transfers between adsorbed molecules and the metallic substrate are involved and the electromagnetic enhancement which considers the excitation of localized surface plasmon (LSP). A detailed explanation about SERS mechanisms can be found elsewhere [16,17]. This technique has the advantage that it makes possible to quench a common interference found in Raman spectra: the fluorescence. In literature various SERS substrates have been proposed, and the use of solid metallic substrates such as electro-polished aluminum foils [18] seems to be a good choice because of its simplicity from a practical point of view.

Previous works have demonstrated the ability of spectroscopic techniques such as UV-Visible spectroscopy [19] and ^1H -Nuclear Magnetic Resonance [20] to detect Sudan dyes in food matrices. In case of Raman spectroscopy, the literature concerning Sudan dyes is very limited, and only one reference is found regarding this dye in a complex food matrix [21].

The aim of the present study was to evaluate the potential of Raman spectroscopy as screening method for detecting the banned Sudan I dye in such food matrices as culinary spices. Firstly, we selected the most appropriate Raman modality (normal Raman, FT-Raman or SERS), and then, we studied two chemometric spectra treatments for background removal and noise correction: a smoothing technique such as Savitzky-Golay with polynomial baseline correction and wavelet transform [8, 22-27]. Finally, an unsupervised exploratory analysis such

as Principal Component Analysis (PCA) was applied to evaluate the effect that both spectra treatments have in the differentiation of the unadulterated samples from the ones adulterated with Sudan I dye.

2. Materials and Methods

2.1 Sample preparation and dataset

For normal and FT-Raman, solid paprika powders unadulterated and adulterated with Sudan I dye were analysed. The adulterated sample was obtained by spiking the paprika with a standard solution of Sudan I which was then heated and stirred magnetically to get rid of the solvent (chloroform analytical grade) completely. This sample was homogenized by ultrasonic treatment for 30 minutes, as was the unadulterated sample so that both samples were subject to the same experimental treatment. The final Sudan dye concentration was 0.01 kg/kg.

For SERS analysis, paprika samples were dissolved in chloroform and extracted with syringe filters. The adulterated samples were obtained by spiking the unadulterated extracts with Sudan I dye in such a way that the final concentration was 0.0036 kg/kg. To obtain the SERS spectra, 60 μ l of each sample was uniformly dropped onto aluminum foils and dried to get rid of solvent completely. The final SERS dataset comprised 40 paprika samples, 20 of which were unadulterated samples (10 hot paprikas and 10 mild paprikas) and 20 adulterated with Sudan I dye.

2.2 SERS substrates

High purity 99.999% aluminum (Al) foils (1cm x 1cm) from Goodfellow Cambridge Ltd. with a thickness of 0.25 mm were used as the SERS substrates. The foils were first cleaned in an ultrasonic bath of acetone and rinsed in deionised water. Then, the dried foils were put in the annealing furnace at 400 °C for 3 hours with a nitrogen inlet to prevent oxidation. The following step was to electropolish the aluminum foils in a 4:1 mixture of ethanol and perchloric acid (HClO_4) at 5 °C. The potentiostatic regime of 20 V was applied for 2 minutes to obtain a smooth mirror-finished surface followed by a rinse with pure ethanol and deionised water.

2.3 Raman measurements

Fourier Transform Raman measurements were obtained from a Thermo Nicolet 5700 FT-IR spectrometer equipped with a FT-Raman module NXR with an InGaAs detector. A 1064 nm radiation from a Nd:YAG laser with a laser power of 250 mW at the sample was used for excitation. A total number of 500 scans were made at a resolution of 2 cm^{-1} .

Normal Raman and SERS measurements were acquired using a Renishaw inVia Reflex Raman confocal microscope (Gloucestershire, UK), equipped with an Ar-ion laser at 514 nm, a He-Ne laser at 633 nm, a diode laser emitting at 785 nm and a Peltier-cooled CCD detector (-70 °C) coupled to a Leica DM-2500 microscope. Calibration was carried out daily by recording the Raman spectrum of an internal Si standard. Rayleigh scattered light was appropriately rejected by using edge-type filters. The diffraction grating (1200 lines/mm) gave the spectral range

100-2000 cm^{-1} with a spectral resolution of 2 cm^{-1} . Spectra were recorded with the accumulation of three scans with a 10 s scan time each one.

We made the final SERS measurements by using a 20x working distance microscope lens and a laser power of 100 mW at the sample. To ensure representative measurements spectra from five different points on the aluminium foil were averaged.

2.4 Theory

2.4.1 Savitzky-Golay Smoothing and Polynomial Baseline Correction

The Savitzky-Golay smoothing method [28] fits an m th degree polynomial to a subset of data within a window of defined size superimposed on the spectrum. This window is moved forward in steps of one data point and in each step the central point of the window is replaced by the value obtained from the polynomial fit.

Polynomial baseline correction [29] estimates the baseline as a polynomial function, which is fitted to specified spectral points outside peaks that are visually selected. The polynomial function is then subtracted from the original spectra.

In both cases, the polynomial function is obtained using the least-squares criterion.

2.4.2. Discrete Wavelet Transform

The fundamentals of wavelet analysis have been described in detail elsewhere [26, 30]. Several families of wavelets are available [26], of which we tested two of the most commonly used: Daubechies and Symmlet. In our case, Symmlet 8 was chosen because of its better performance. The use of wavelet analysis for denoising and background removal involves dividing the original signal into low-frequency parts (named approximation components A) and high-frequency parts (named detail components D) according to Mallat's pyramidal algorithm [31]. The approximation component is again divided according to the above criterion until the j level is reached, in such a way that the original signal (s) is expressed as $s = A_j + D_j + D_{j-1} + \dots + D_1$. As an example, Figure 1 shows the decomposition of a Raman signal in 7 levels. It can be seen that noise is contained mainly in the first detail components while the smooth part of the signal is contained in the approximation component (A7). So, when the approximation component at level j does not contain a significant Raman signal, it can be set as the background signal and removed in the reconstruction step. Similarly, noise appears in the first detail components (D1 and D2, Figure 1), so to remove it from the spectra, soft thresholding is applied. A threshold value is calculated on the basis of Stein's Unbiased Estimate of Risk [32] and the detailed components with values lower than this threshold are set to zero and new detail components (D1_T and D2_T) are obtained (Figure 1). Finally the spectrum is reconstructed using these new detail components and subtracting the A7 approximation component (Figure 1), thus obtaining the transformed signal expressed as $s_{if} = D1_T + D2_T + D3 + \dots + D7$.

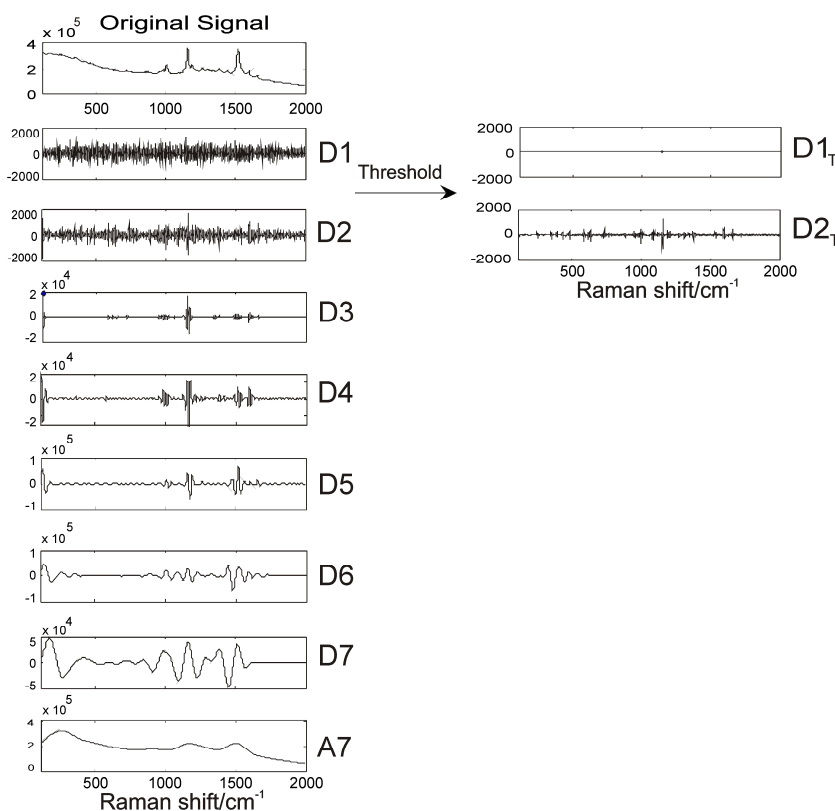


Figure 1. Wavelet decomposition of a Raman signal at level 7. A7 is the approximation component while D1 to D7 are the detail components, where D1_T and D2_T are the components after thresholding has been applied.

2.4.3 Principal Component Analysis (PCA)

PCA is a well-known unsupervised technique that projects high dimensional data onto lower dimensional space [33]. The first principal component (PC1) is determined by retaining the maximum variance (maximum information). Each subsequent principal component describes a maximum of variance that is not modeled by the former components, so most of the variance is contained in the first

few principal components (PCs). All redundant information is summarized, thus simplifying the graphical interpretation of the data.

3. Results and Discussion

3.1. Preliminary studies

Samples were firstly studied by normal Raman at three laser wavelengths (514, 633 and 785 nm) and Raman spectra of poor quality were obtained. As an example, spectra obtained at 785 nm for solid powder spice unadulterated (i) and adulterated with Sudan I (ii) are shown in Figure 2.a. Both spectra can be observed to lack signals, which are probably masked by the high levels of fluorescence that they present.

Otherwise, when the spectra for the two samples mentioned above were recorded in an FT-Raman operating with a 1064 nm laser (Figure 2.b), the fluorescence still has an effect and the signal-to-noise ratio is poor, which does not allow visual differences to be found between the unadulterated and adulterated spectra as in the above case. The results at this point, then, show that the fluorescence given off by the complex matrix interferes quite considerably and masks the weak Raman signal.

Finally, to analyse the samples by the SERS technique, they were deposited on aluminium foils and the spectra were obtained with three different wavelength lasers (514, 633 and 785 nm). As an example, Figure 3 shows the spectra for the same unadulterated and adulterated spices shown in the previous figure. It can be seen that at 514 nm, the unadulterated (spectrum a) and adulterated samples (spectrum b) give spectra which have poor signals and, in the latter case with a large

baseline drift. Moreover, at 633 nm, the signal obtained from the unadulterated (spectrum a) and adulterated samples (spectrum b) cannot be considered a proper Raman signal. In both cases, the results of varying the Raman conditions (laser power and lens) were not satisfactory. On the other hand, at 785 nm (Figure 3) a characteristic Raman spectrum can be obtained. In this case, the adulterated sample (spectrum b) presents peaks belonging to the adulterant dye which are not present in the unadulterated sample (spectrum a). Considering all these results, we will now go on discussing the most appropriate chemometric approach working with the SERS technique with the laser at 785 nm.

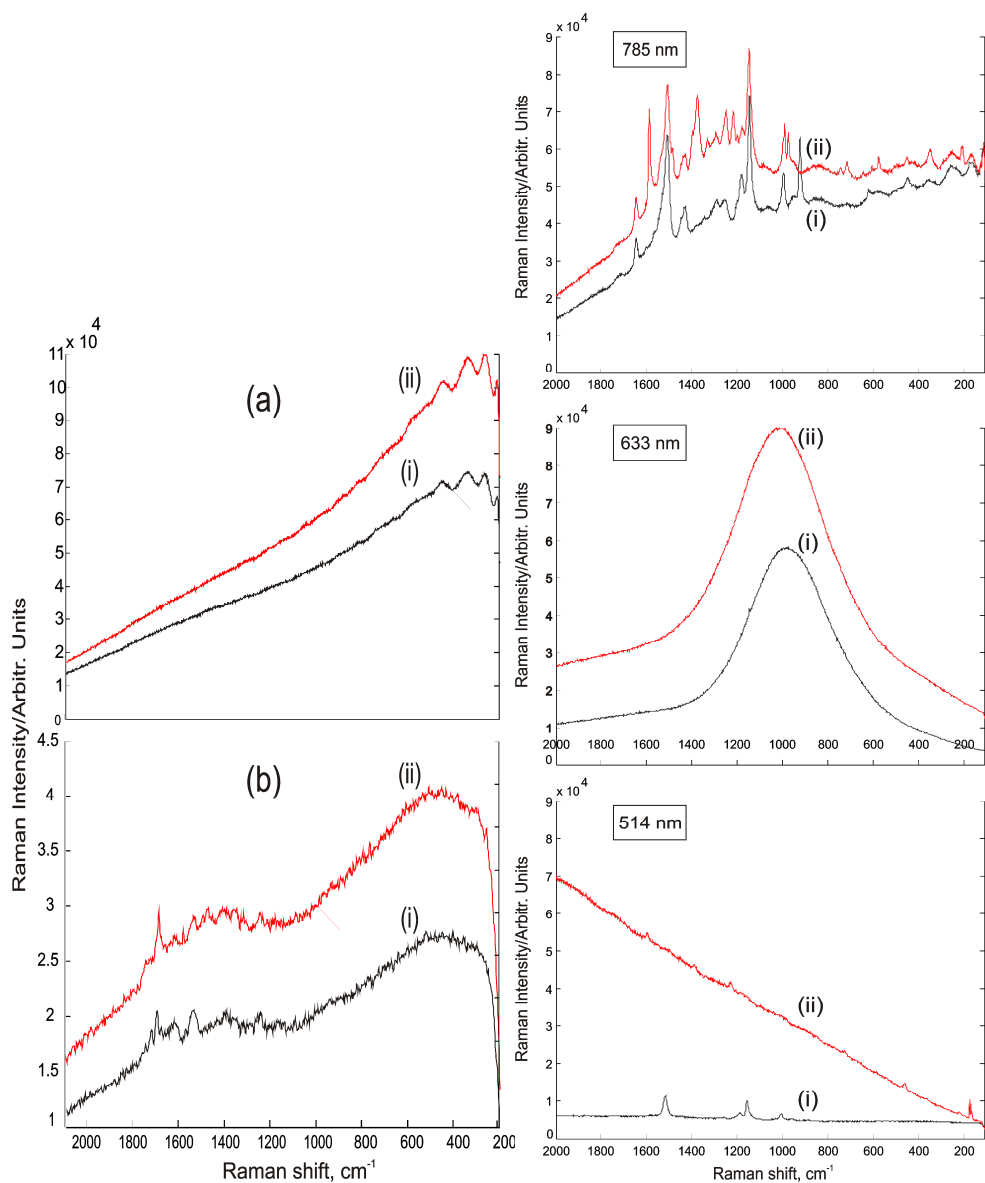


Figure 2 (left). Raman spectra of solid paprika samples measured with (a) normal Raman and (b) FT-Raman. In both cases, an (i) unadulterated sample and (ii) sample adulterated with Sudan I are shown.

Figure 3 (right). SERS spectra of (i) an unadulterated sample and (ii) the same adulterated sample measured with different wavelength lasers: 514, 633 and 785

nm.

Figure 4 shows as an example, the SERS spectra of (i) a random unadulterated paprika spice, (ii) the same spice adulterated with Sudan I, (iii) the pure Sudan I dye and (iv) the SERS metallic substrate. The spectrum of the substrate presents a broad band around 1350 cm^{-1} which causes a baseline drift in the last part of the spectrum. The most intense peaks both for the pure dye and for the unadulterated sample, appear between 900 and 1650 cm^{-1} , but just by a visual inspection it is difficult to assure whether a sample contain or do not contain the adulterant Sudan I dye. For example, by examining the Raman zone between 100 and 600 cm^{-1} in the spectrum of the adulterated sample, some peaks from the Sudan I dye can be masked by the food matrix signal, the background effect and/or the lower concentration level. Also, it can be mentioned that there are some peaks belonging to the food matrix around 1000 , 1160 and 1520 cm^{-1} (see Figure 4.i) which have a considerable intensity that completely overlap the peaks from the Sudan dye. However, some characteristic bands of the Sudan I dye (Figure 4.iii) can be seen in the spectrum of the adulterated sample (shown in Figure 4.ii with arrows) which are not present in the spectrum of the unadulterated sample (Figure 4.i). This situation can be considered a particular case as it is highly depend on the sample/matrix composition and therefore can not be extended to future samples. In view of the aforementioned, the use of multivariate tools to process Raman spectra constitutes an advantageous way to extract the most important and relevant information when samples suspicious to be adulterated want to be recognized.

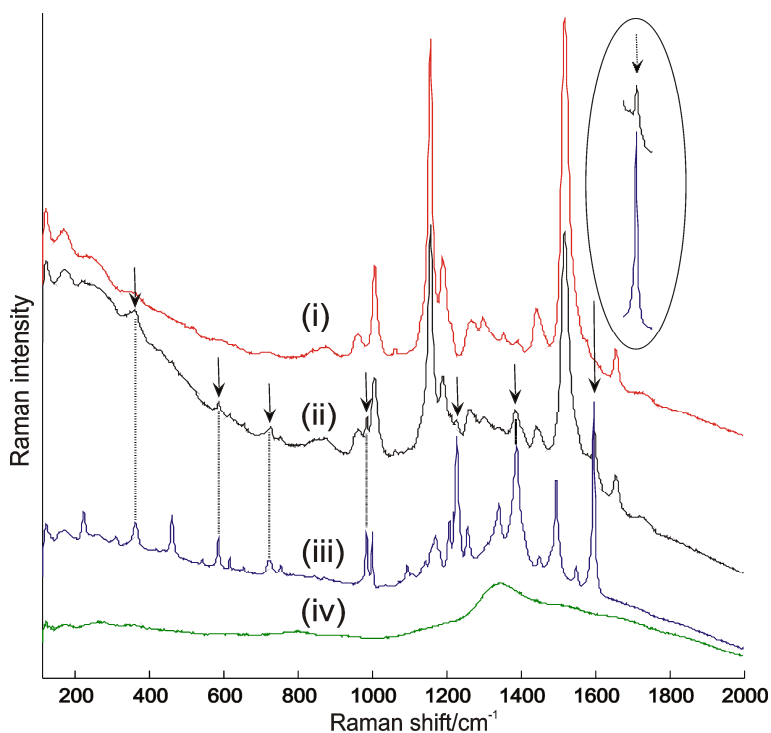


Figure 4. SERS spectra at 785 nm of (i) an unadulterated sample, (ii) an adulterated sample, (iii) the pure Sudan I dye, and (iv) the metallic substrate.

3.2. Spectra treatment

For the Savitzky-Golay smoothing, two parameters have to be fixed: the size of the window and the order of the polynomial. Results were best with a five point window and a first-order polynomial. To adjust the baseline by means of a polynomial function, 20 representative points were selected. Figure 5.a shows, as an example, a raw adulterated sample in which the selected points are depicted with arrows together with the polynomial function, which in this case, a third order function was the optimal choice. Figure 5.b shows the smoothed and baseline-corrected spectrum and it can be seen that, after the treatment, the spectrum retains the original shape of the Raman peaks and the background has been removed.

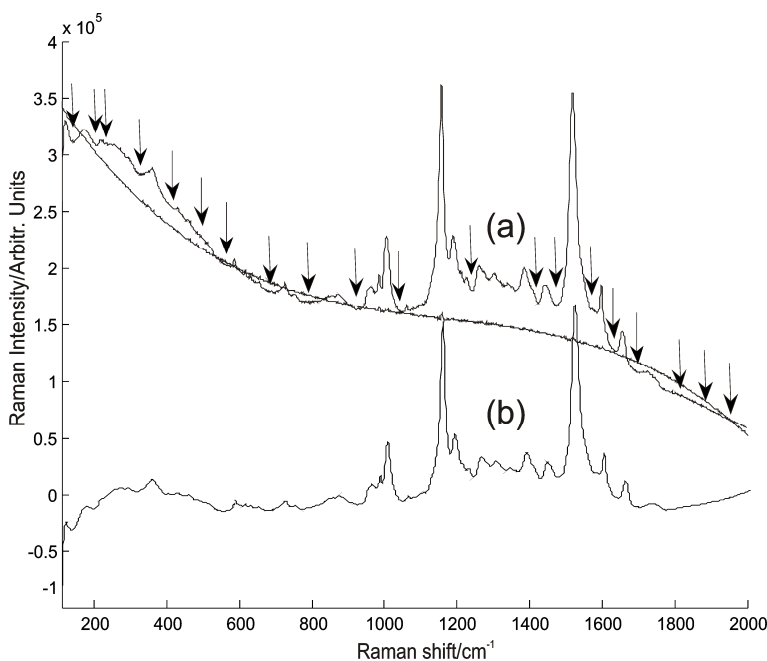


Figure 5. (a) Raw spectrum with the selected points representative of the baseline marked with arrows and (b) spectrum after Savitzky-Golay smoothing and polynomial baseline correction treatment.

For the wavelet transform, the optimal decomposition level was chosen by a visual inspection of the detail and approximation components (Figure 1). In our case the optimal decomposition level was the seventh, because at lower decomposition levels the approximation coefficients still contain some Raman signals, while at upper levels the reconstructed spectra does not significantly change. Figure 6.a shows the raw spectra (same sample as in the previous figure) and the spectra after the application of the wavelet transform (Figure 6.b). As in the other spectra treatment it can be seen that the Raman peaks are well conserved after the denoising process. In general, the background was efficiently removed but the first part of the spectrum (between 100 and 300 cm^{-1}) still retains some background and

in the zone between 1100 and 1600 cm^{-1} the removal was a little more pronounced but the Raman peaks were not significantly altered.

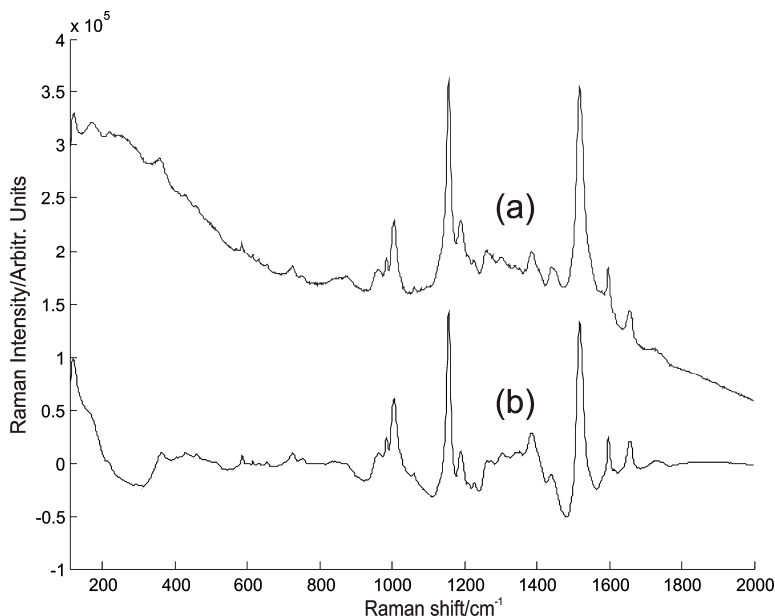
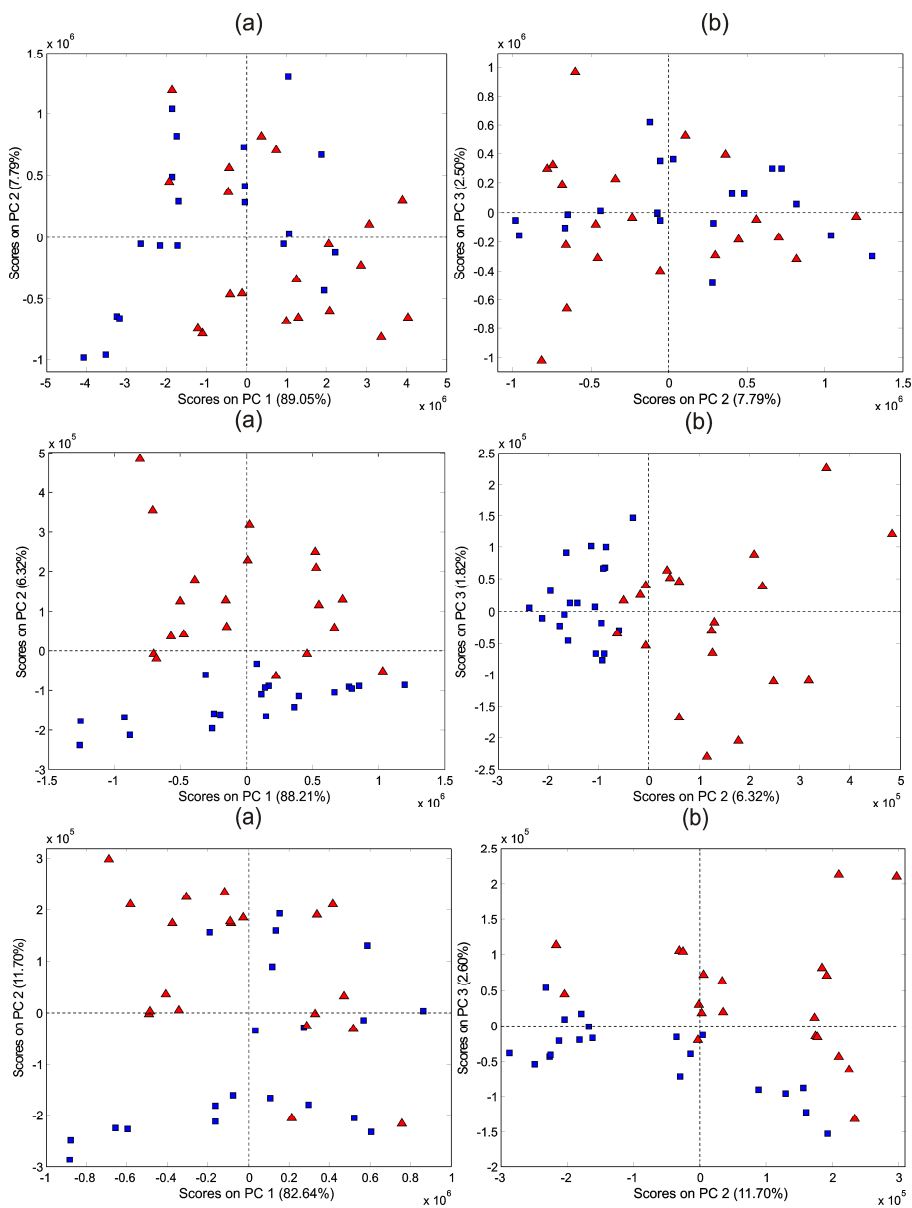


Figure 6. (a) Raw SERS spectrum and (b) the corresponding spectrum after the application of the wavelet transform.

3.3. Exploratory Analysis

PCA was applied as an exploratory analysis to visualize the distribution of the samples before and after the spectra treatment. Data was mean-centered and PCA was applied both to raw Raman spectra and to the spectra treated by the two different pre-treatment strategies. Figures 7, 8 and 9 shows the PCA scores plots for (a) PC1 vs. PC2 and (b) PC2 vs. PC3 for untreated data (raw data), data treated with the Savitzky-Golay and polynomial baseline fitting and data treated with wavelet transform, respectively. In all cases, the first PC retains a high percentage of the total variance (89.0, 88.2 and 82.6, respectively) and with this

information the adulterated samples (marked as squares) cannot be distinguished from the unadulterated ones (marked as triangles). This indicates that most information is related to the sample composition and not to the adulterant. However, the greatest differences between both groups of samples could be seen from data processed with the two strategies when using the information provided by PC2 and PC3. In case of Savitzky-Golay and polynomial baseline correction (Figures 8.a and 8.b) treatment, the separation is achieved mainly along PC2 in which unadulterated samples have the most positive scores and the adulterated samples the most negative scores. On the other hand, regarding data processed with wavelet transform, PC3 plays the most important role in the discrimination (Figure 9.b) and both groups of samples follow a similar distribution (positives and negatives scores) along PC3 as in the previous case. The discrimination among the different samples is not possible when dealing with raw data (Figure 7).



Figures 7, 8 and 9. PCA scores plots for untreated data (upper), data processed with Savitzky-Golay and polynomial baseline correction (medium) and data processed with wavelet analysis (down), respectively. In all the cases, (a) PC1 vs. PC2 and (b) PC2 vs. PC3 are depicted.

4. Conclusions

This work was focused on developing tools for a rapid and reliable detection of Sudan I dye added illegally to culinary spices by means of normal Raman, FT-Raman and SERS. SERS is the most appropriate modality capable of providing a proper Raman signal when a complex matrix such as paprika spice is analyzed.

The advantage of processing Raman spectra before the application of multivariate analysis is demonstrated. The application of PCA to raw data does not show different groups between unadulterated and adulterated samples, but when applied to data processed either with Savitzky-Golay smoothing combined with polynomial baseline correction or wavelet transform, the two groups of samples are differentiated. These results reveal that both strategies are suitable for processing Raman spectra as they give similar results.

The results obtained by PCA show that although PC1 contains most of the total information, it is unable to separate unadulterated and adulterated samples. In contrast, the separation becomes possible when information about PC2 and PC3 is used. The preliminary results provided by an unsupervised exploratory analysis suggest that this methodology can potentially be used for classification purposes in a future research.

The proposed methodology shows that SERS combined with appropriate spectra treatment can be used as a practical screening tool to distinguish samples suspicious to be adulterated with Sudan I dye. The recent introduction of portable Raman systems opens the way to implement SERS as a rapid method to achieve on-site detection, with the advantage to use simple SERS substrates that do not require many experimental difficulties. In a future research, different metallic substrates can be evaluated to optimize SERS analysis, and this methodology can be

extended to other complex food matrices subjected to be adulterated such as turmeric, curry or chilli powder.

Acknowledgments

Carolina Di Anibal thanks the Agency for the Administration of University and Research Grants of the Catalan Government (AGAUR) for providing the doctoral fellowship. Lluís F. Marsal thanks the Spanish Ministry of Science and Innovation (MICINN) under grant number TEC2009-09551.

5. References

- [1] M. Stiborová, V. Martínek, H. Rýdlová, P. Hodek, E. Frei, *Cancer Res.* **62** (2002) 5678-5684.
- [2] Commission decision of 23 May 2005 on emergency measurements regarding chilli products, curcuma and palm oil, *Official Journal of the European Union* (2005/402/EC) L135/34.
- [3] R. Rebane, I. Leito, S. Yurchenko, K. Herodes, *J. Chromatogr. A* **1217** (2010) 2747.
- [4] E.G. Anastasaki, C.D. Kanakis, C. Pappas, L. Maggi, A. Zalacain, M. Carmona, G.L. Alonso, M.G. Polissiou, *J. Agric. Food Chem.* **58** (2010) 6011.
- [5] D.I. Ellis, D. Broadhurst, S.J. Clarke, R. Goodacre, *Analyst* **130** (2005) 1648.
- [6] R.C. Barthus, R.J. Poppi, *Vib. Spectrosc.* **26** (2001) 99.
- [7] Y. Kim, S. Lee, H. Chung, H. Choi, K. Cha, *J. Raman Spectrosc.* **40** (2009) 191.
- [8] A.M. Nikbakht, T. Tavakkoli Hashjin, R. Malekfar, B. Gobadian, *J. Agric. Sci. Technol.* **13** (2011) 517.
- [9] R.M. El-Abassy, P.J. Eravuchira, P. Donfack, B. von der Kammer, A. Materny, *Vib. Spectrosc.* **56** (2011) 3.
- [10] C. Fan, Z. Hu, L.K. Riley, G.A. Purdy, A. Mustapha, M. Lin, *J. Food Sci.* **75** (2010) M302.
- [11] X.F. Zhang, M.Q. Zou, X.H. Qi, F. Liu, X.H. Zhu, B.H. Zhao, *J. Raman Spectrosc.* **41** (2010) 1655.
- [12] B.S. Luo, M.Lin, *J. Rapid Methods Autom. Microbiol.* **16** (2008) 238.
- [13] L. He, N.J. Kim, H. Li, Z. Hu, M. Lin, *J. Agric. Food Chem.* **56** (2008) 9843.
- [14] M. Lin, L. He, J. Awika, L. Yang, D.R. Ledoux, H. Li, A. Mustapha, *J. Food Sci.* **73** (2008) T129.
- [15] Y. Cheng, Y. Dong, *Food Control* **22** (2011) 685.
- [16] E.C. Le Ru, E. Blackie, M. Meyer, P.G. Etchegoin, *J. Phys. Chem. C* **111** (2007) 13794.
- [17] C.L. Haynes, A.D. McFarland, R.P. Van Duyne, *Anal. Chem.* **77** (2005) 338A.
- [18] X. Zhou, Y. Fang, P. Zhang, *Spectrochim. Acta Part A* **67** (2007) 122.
- [19] C.V. Di Anibal, M. Odena, I. Ruisánchez, M.P. Callao, *Talanta* **79** (2009) 887.
- [20] C.V. Di Anibal, I. Ruisánchez, M.P. Callao, *Food Chem.* **124** (2011) 1139.
- [21] W. Cheung, I.T. Shadi, Y. Xu, R. Goodacre, *J. Phys. Chem. C* **114** (2010) 7285.

- [22] A. Kwiatkowski, M. Gnyba, J. Smulko, P. Wierzba, *Metrol. Meas. Syst.* **17** (2010) 549.
- [23] J.F. Brennan, Y. Wang, R.R. Dasari, M.S. Feld, *Appl. Spectrosc.* **51**(1997) 201.
- [24] Y.Y. Huang, C.M. Beal, W.W. Cai, R.S. Ruoff, E.M. Terentjev, *Biotechnol. Bioeng.* **105** (2010) 889.
- [25] J. Li, L.P. Choo-Smith, Z. Tang, M.G. Sowa, *J. Raman Spectrosc.* **42** (2010) 580.
- [26] P. Ramos, I. Ruisánchez, *J. Raman Spectrosc.* **36** (2005) 848.
- [27] Y. Hu, T. Jiang, A. Shen, W. Li, X. Wang, J. Hu, *Chemom. Intell. Lab. Syst.* **85** (2007) 94.
- [28] A. Savitzky, M.J.E. Golay, *Anal. Chem.* **36** (1964) 1627.
- [29] T.J. Vickers, R.E. Wambles, C.K. Mann, *Appl. Spectrosc.* **55** (2001) 389.
- [30] B. Walczak, D.L. Massart, *Chemom. Intell. Lab. Syst.* **36** (1997) 81.
- [31] S.G. Mallat, *IEEE Trans. Pattern Anal. Mach. Intell.* **11** (1989) 674.
- [32] D.L. Donoho, L.M. Johnstone, *J. Am. Stat. Assoc.* **90** (1995) 1200.
- [33] M. Otto, *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, Wiley-VCH Weinheim, Germany (1999).

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

3.2. Implementation of chemometric strategies to improve classification results

This section contains three papers that focus on using certain chemometric tools to improve the performance of a classification model.

The first paper (Talanta 86 (2011) 316-323) addresses the importance of selecting variables when working with large datasets. The paper presents a case study in which NMR data are evaluated, and compares three variable selection methods: Xdiff, interval-PLS (iPLS) and Genetic algorithms (Gas). The idea is to evaluate the performance of such methods to extract the most relevant variables prior to the application of a classification technique, which in our case is PLS-DA. The different variable selection methods are compared by evaluating the classification results in the same way as in previous papers.

The second paper (Talanta 84 (2011) 829-833) assesses the potential of combining data from different sources. Given that it is now common to have more than one instrument in a laboratory, more information regarding a particular analytical problem can be obtained. Therefore the synergy between UV-Visible and $^1\text{H-NMR}$ data has been evaluated by means of two data fusion strategies: variable and decision level. The idea is to improve the classification results obtained when evaluating the spectroscopic techniques individually. The performance of each data fusion strategy is evaluated and compared using their classification ability; that is, they are evaluated and compared in terms of the types of PLS-DA classification error they make, as in previous papers.

The third paper (Submitted for publication) evaluates the feasibility of applying multivariate calibration transfer methods (standardization methods) to a classification framework. The idea is to maintain the classification model's ability

for to predict future samples over given period of time. Sometimes, the spectra measured in different conditions (new samples) differ from the spectra used to build the initial model with the result that these new samples might be wrongly predicted. Consequently, applying standardization methods is a useful strategy for solving this problem. This thesis uses Piecewise Direct Standardization (PDS) to correct UV-Visible spectra from different experimental conditions, in order to fit these spectra to the initial conditions. In doing so, the benefits and usefulness of the standardization process are highlighted.

3.3.1. PAPER

¹H-NMR Variable Selection Approaches for Classification. A Case Study: The Determination of Adulterated Foodstuffs

Carolina V. Di Anibal, M. Pilar Callao, Itziar Ruisánchez

Talanta 86 (2011) 316-323

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

$^1\text{H-NMR}$ Variable Selection Approaches for Classification. A Case Study: The Determination of Adulterated Foodstuffs

Carolina V. Di Anibal, M. Pilar Callao, Itziar Ruisánchez

¹Chemometrics, Qualimetrics and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n Tarragona, Spain.

Abstract

Whenever dealing with large amount of data as is the case of a NMR spectrum, carrying out a variable selection before applying a multivariate technique is necessary. This work applies to various variable selection techniques to extract relevant information from $^1\text{H-NMR}$ spectral data. Three approaches have been chosen, because each is based on very different foundations. The first method, called Xdiff, is based on calculating the normalized differences between the mean spectrum of a class considered to be the reference and the spectra of each sample. The second approach is the interval Partial Least Squares method (iPLS), which investigates the influential zones of the spectra that contains the most discriminating predictors calculating local PLS-DA models on narrow intervals. The last one is Genetic Algorithms (GAs) which finds the optimal variables from a random initial subset of variables by means of an iterative process. The performance of each variable selection strategy is determined by the classification results obtained when multiclass Partial Least Squares-Discriminant Analysis is applied. This study has been applied to NMR spectra of culinary spices that might be

adulterated with banned dyes such as Sudan dyes (I-IV). The three techniques give neither the same number nor the same selected variables, but they do select a common zone from the spectra containing the most discriminating variables. All three techniques give satisfactory classification and prediction results, being higher than 95% with iPLS and GA and around 89% with Xdiff, therefore the three variable selection techniques are suitable to be used with NMR data in the determination of food adulteration with Sudan dyes as well as the specific type of adulterant used (I-IV).

Keywords: Variable selection, Nuclear Magnetic Resonance, Food adulteration, Sudan Dyes, PLS-DA.

1. Introduction

In food industry, the addition of some colorants to some products is a common practice, as colorants enhance its visual aesthetics and promote sales. Up to now four Sudan (I-IV) dyes have been detected in certain food products, as culinary spices destined to human consumption, although they are normally used for colouring plastics and other synthetic materials. The current European legal framework on colours in food establishes that Sudan dyes are not included in the list of authorized colorants, as these dyes have potential carcinogenic effects [1] and even more, Sudan I may also have genotoxic effects. Therefore Sudan dyes are banned to be used as additive in food matrices for human consumption.

Previous studies have demonstrated that Proton Nuclear Magnetic Resonance Spectroscopy ($^1\text{H-NMR}$) is a well suited analytical technique capable of detecting these four Sudan dyes when they are as adulterants in culinary spices [2]. NMR is a technique that generates a specific profile of the sample studied. Recent

breakthroughs in NMR technology have led to measurements with increased sensitivity, resolution and reproducibility, thereby contributing to the production of high quality data [3]. These improvements in data quality, coupled with multivariate techniques, have given rise to a well-known rapid screening method [4] which has been demonstrated to be an efficient method for food screening, discrimination and characterization [2, 5-7].

Because of the large amount of data obtained from a $^1\text{H-NMR}$ spectrum, carrying out a variable selection before applying a multivariate technique is common practice. There are many potential benefits of variable selection: facilitating data visualization and data understanding, reducing the variables/samples ratio, eliminating noisy variables as well as redundant information, among others. All these advantages are important when chemometrics methods want to be applied. Many methods are found in the bibliography for variable selection in NMR in classification problems. Some methods include reducing noisy variables or variables of low intensity or even bucketing, with the consequent reduction of data. Other methods are supervised with the aim to find which variables are the most discriminatory in order to achieve the best discrimination when working with different groups of samples or classes. Some examples of both mentioned types found in literature include stepwise discriminant analysis [8, 9], supervised variable selection methods [10], self organizing maps (SOMs) [11], PLS weight coefficients [12], wavelet transform [13], univariate selection based on the maximum intensity differences [14], interval Partial Least Squares (iPLS) [15, 16] and Genetic Algorithms (GAs) [17, 18]. This wide variety of variable selection techniques implies that choosing the most appropriate one for a specific problem it is not an easy task.

The aim of this work is to study the ability of three supervised techniques for selecting variables when working with NMR spectral data in the Sudan dyes

classification problem, as it is well known that choosing the optimal variable selection technique is problem dependent. The three approaches studied have been chosen because each is based on a very different principle. Our work focuses on the region and number of selected variables, the number and type of misclassified samples and the overall performance of the classification process obtained with each technique.

The first technique is based on the computation of new variables called X_{diff} , which are the normalized differences between the mean spectrum of a class considered to be the reference and the spectra of each sample. The hypothesis is that variables having different intensities will be reinforced in the new X_{diff} values, thus enabling differentiation between the classes. The second approach is the interval Partial Least Squares method (iPLS) [19], which investigates the influential zones of the spectra that contains the most discriminating predictors, and calculates local PLS-DA models in narrow intervals. The third variable selection technique used is the well known Genetic Algorithms (GAs) [20, 21], which can find the optimal variables from a random initial subset of variables by means of an iterative process. The performance of each variable selection strategy is determined from the classification results obtained when Partial Least Squares-Discriminant Analysis is applied.

2. Data Analysis Methods

2.1. Variable selection

2.1.1. Xdiff method

This variable selection method [2, 14] is applied to a multiclass problem and it is based on calculating the x_{diff} values in accordance with Eq. (1)

$$x_{diff,ij} = \frac{|x_{ij} - \bar{x}_i|}{\sigma_i} \quad (1)$$

where x_{ij} is the i^{th} variable for the j^{th} sample and \bar{x}_i and σ_i are the mean and standard deviation, respectively, calculated from each i^{th} variable obtained from a reference class. As we are dealing with an adulteration problem, among our predefined five classes, we have set the unadulterated class as the reference one.

The **Xdiff** matrix is calculated for all five classes. A threshold value was defined from the x_{diff} values of the reference class through a visual inspection in a way that most of the x_{diff} values are kept below. Therefore, only those original variables that correspond to x_{diff} values higher than the prefixed threshold are selected. Several threshold values around the first prefixed one are checked, retaining the one which gives the best PLS-DA classification results.

2.1.2. Interval Partial Least Squares method

Interval PLS (iPLS) develops local PLS-DA models on equidistant subintervals of the full-spectrum region and the prediction performance of these

local models and the global (full-spectrum) model is compared, mainly by means of the validation parameter RMSECV (root mean squared error of cross-validation, (Eq. 2)):

$$\text{RMSECV} = \frac{\sqrt{(\sum \hat{y}_i - y_i)^2}}{n} \quad i = 1, \dots, n \quad (2)$$

with y_i as the true class assignment value for sample i , \hat{y}_i as the predicted class assignment value from cross-validation and n as the number of samples.

iPLS provides an overall picture of the relevant information in different spectral subdivisions, thereby removing non-relevant information from other regions.

2.1.3. Genetic Algorithms (GAs) method

The GA theory is explained in detail elsewhere [20, 21], so we will limit ourselves to show the procedure followed in the present study depicted in Figure 1 which involves an iterative process. As for running the GA algorithm, the adequate number of input variables has to be not far away from 200, the original NMR variables have to be reduced, so the average of “ n ” consecutive variables is obtained. To decide the optimal window size n , the Principal Components Analysis (PCA) score plots of both, the original and the mean of n variables, are compared until a similar distribution of samples is kept. The next step is to apply the GA algorithm to the mean variables, with a previous step in which a randomization test is applied to check whether the dataset is adequate to run the algorithm, in order to avoid the overfitting problem commonly found in a GA-based feature selection [22]. Therefore, the mean variables selected by GA are expanded to the n consecutive

original ones used to obtain the mean variables. Finally, PLS-DA is applied to those original ones.

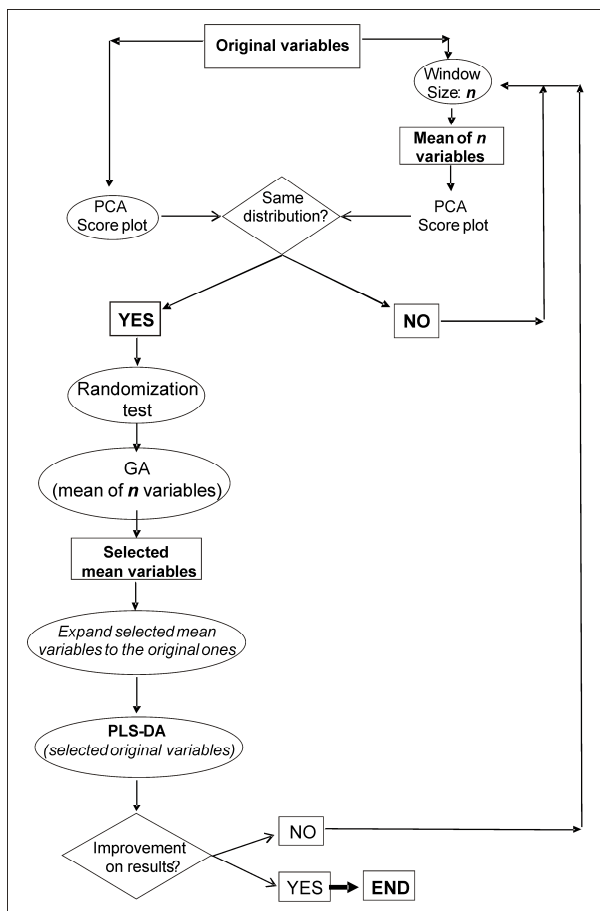


Figure 1. Genetic Algorithm scheme for the iterative variable selection process.

All this procedure is done iteratively until a subset of variables giving the optimal classification results comparing the previous and last iteration is found. It has to be remarked that the second (and so on) iterations start with the previous subset of “expanded original variables”, so a lower n value is looked for.

2.2. Classification method: Partial Least Squares-Discriminant Analysis

PLS-DA is a regression technique adapted to a supervised classification task [23]. A PLS regression model is calculated, which relates the independent variables (e.g. spectra) to a binary “y” vector which has as many values as classes in order to designate the class of the sample. For example, a vector [1,0,0,0,0] means that of five possible classes, the sample belongs to class 1, and so on. Classification of an unknown sample is derived from the value predicted by the PLS model, \hat{y} . Ideally, this value should be close to the values used to codify the class (here either 0 or 1). A threshold value for each pre-defined class is defined between 0 and 1 so that a sample is assigned to the class for which its prediction is larger than the threshold value. Typically, a normal distribution fits \hat{y} values and the threshold value is estimated using Bayes’ rule [24]. The optimal number of latent variables (LVs) was chosen to minimize the root mean square-cross validation prediction error (RMSECV) for all the classes. This number is selected through a compromise between the optimal values for each class.

Both, class assignment and the percentage of predicted probability are considered for the evaluation of the multiclass PLS-DA classification results. In this study, the selected variables are auto-scaled before running the multiclass PLS-DA algorithm.

2.3. Training and test set

In order to avoid overoptimistic results, data set is to divide into training and test set, using the training set to select the most relevant variables [25]. In our case, approximately 15% of samples of each class are left out to form the test set. Principal Component Analysis (PCA) scores plots (not shown) are used to select

the representative samples that form the test set by using two different strategies: a selection based on a randomly choice and on Principal Component Analysis (PCA) scores plots. In the last case, the representative samples are selected in a way to cover each class spatial distribution. Both strategies give similar global classification results, so we decided to use PCA because it selects the samples in a more homogeneous manner. Finally, the training set and test set are used to validate the PLS-DA models.

3. Experimental and data

3.1. Samples and NMR data

The studied samples correspond to twenty seven unadulterated commercial spices previously checked by HPLC-DAD they are free of any Sudan dye [26]. The samples are prepared weighing 0.1 g of each spice, dissolving it in 5 mL of deuterated chloroform and filtered. From this solution, 700 μL were taken and placed in 2 mL flasks. Spiked samples were prepared by adding the stock Sudan (I–IV) solutions to each commercial sample to obtain a concentration within a range in which adulterated spices are commonly found: 7.1 g kg^{-1} [27]. So, five groups of twenty seven samples each give a total of 135 samples, where class 1 corresponds to the unadulterated samples and classes 2-5 correspond to the adulterated samples with Sudan I-IV, respectively.

The $^1\text{H-NMR}$ spectra was acquired in Varian NMR System 400 at 400.13 MHz using a 4 μs pulse (45°), an acquisition time of 2.2 s (32,768 complex points) with a 15 s delay time to allow full relaxation and a spectral width of 7217 Hz (18 ppm). Sixteen scans were recorded per sample. Spectra were processed using Mestrec-C version 4.7.0 software. All the free induction decays

(FID) were Fourier transformed (FT) by applying, first, a 32 k zero filling and, then, an exponential filter function with line broadening (LB) of 0.5 Hz. The spectra were automatically phase corrected and the baseline correction was made by manual multipoint with 14 points interpolated by a cubic spline function. All spectra were calibrated by setting the CDCl_3 peak at 7.26 ppm. The spectral region between 0.5 and 8.9 ppm is selected, because it is the zone where most of relevant signals are located, although the entire spectrum was used to make baseline corrections. The most intense range signal coming from the solvent was eliminated, leading to 8391 final variables. These variables constitute the raw spectra used for the different variable selection processes.

3.2. Software

All algorithms were run in the Matlab 6.5 (The MathWorks, Natick, MA) computing environment and PLS Toolbox 3.5 (Eigenvector Research Incorporated). The Matlab packages for the iPLS Toolbox as well as the PLS-GA Toolbox are freely available [28].

4. Results and Discussion

As an initial trial, PLS-DA was applied without selecting any variables, being the recognition and prediction ability around 50%. These results suggest that a variable selection is quite necessary. Therefore as a preliminary study, the aromatic zone (from 6.6 to 8.9 ppm) was selected just based on a visual inspection of the spectra, as Sudan dyes have most of NMR signals there, while samples without Sudan dyes do not have any relevant signals. The percentage of correct classification

was 87.8% and 85% for the training and test set, respectively. These preliminary results, although they are quite satisfactory, might be improved if a systematic variable selection approach is implemented.

Regarding the Xdiff approach, as a result of applying Eq. (1), the x_{diff} maximum values are around 40 while the maximum x_{diff} value obtained for the non-contaminated samples is 7 times lower. To select the optimal threshold value, several values were checked on the basis of the minimum classification error obtained. The optimal value of 5.5 is resulted in 1059 selected variables. These selected variables are in accordance with previous studies [2].

For the iPLS variable selection, the raw spectra are first divided into 20 equally sized subintervals containing 420 variables each, and the PLS-DA models were applied to each one. Figure 2.a shows the RMSECV results for each NMR interval, and for the global model (shown by the dotted line). At that point, the strategy is to select the intervals with an RMSECV value below the dotted line (intervals 2-4, 7 and 15) and to make a second iPLS selection. A further division into 10 equally sized subintervals is made (Figure 2.b) and intervals 4 and 5 are those that are located under the RMSECV global value. The classification results considering the 420 variables included in those intervals improve the previous classification results so they are kept as the final selected variables.

Among the different GA structures that can be used, we selected the GA parameters that have been successfully applied to spectral datasets [29]. The optimized GA parameters are those as follows: population size, 30 chromosomes; probability of mutation, 1%; probability of cross-over, 50%; cross-validation, 5 deletion groups; number of runs, 100; maximum number of features selected in the same chromosome, 30; average number of features per chromosome in the original population, 5.

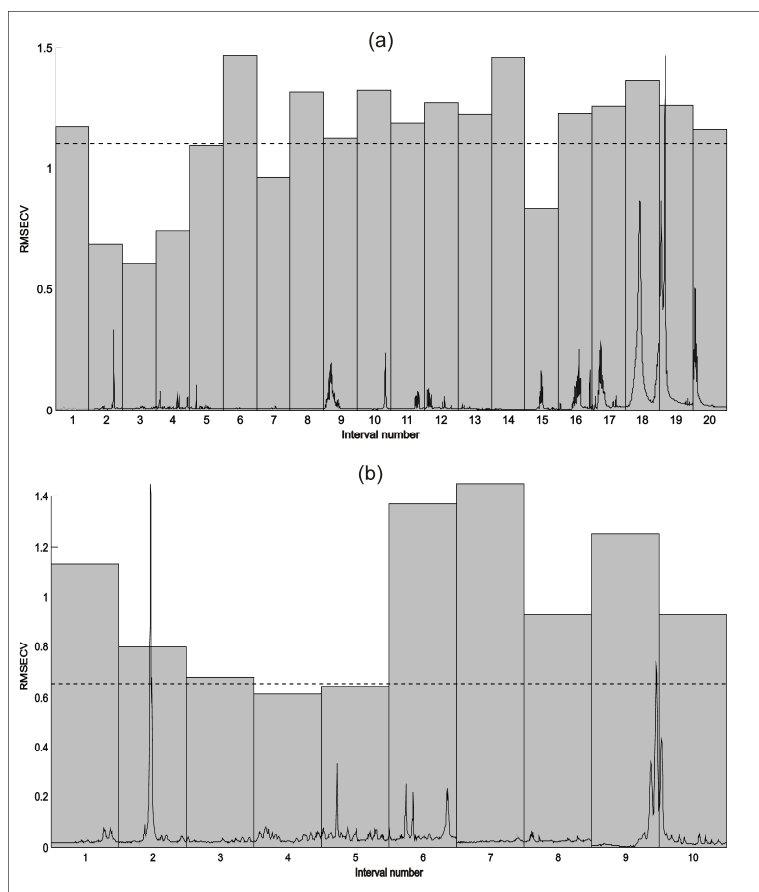


Figure 2. iPLS plots of the cross-validated classification performance (RMSECV) (a) with the first 20 intervals and (b) with the last ten intervals obtained from the previous selected ones. In both plots, the dash line corresponds to the RMSECV value for the global model.

As mentioned above, Figure 1 shows the scheme for the GA iterative feature selection process used in this study. The first window size “ n ” is fixed to a value of 25. This value is chosen looking at the PCA scores plots (not shown) of the original variables and the mean centered variables in order to assure that a similar distribution of samples is maintained. With $n = 25$, 336 mean variables are obtained which corresponds to 8391 divided by 25. Then GA selects 96 mean

variables that when there are expanded to the original ones, correspond to 2400 (96 x 25) original variables. In the second iteration, the maximum value for n which gives comparable PCA scores distribution was 5. Then GA is applied over 480 mean variables selecting 141, which when are expanded correspond a final number of 705 original variables (141 x 5). Further iteration trials do not improve the PLS-DA classification results.

As the final number of selected variables is not fixed in advance, each technique selects a different number: 1059, 705 and 420 for Xdiff, GA and iPLS, respectively. In order to compare the three variable selection methods, Figure 3 shows the original and selected variables obtained with the three selection methods studied for a random adulterated sample. Also, both the spectrum of the pure dye contained in that sample and the unadulterated original sample are shown. Table 1 presents the variables which are most discriminant in each PLS-DA model by evaluating the loading weights of the two first PLS components. The first column presents the NMR chemical shift for the mentioned variables, the second column shows the correspondence with the Sudan dye signals and the last three columns contain the first eight selected variables by each method.

As it can be seen from Figure 3, the spectrum of the pure Sudan II dye has most of its signals in the aromatic region between 6.8 and 9 ppm (Figure 3.b) and some in the aliphatic zone between 1 and 3 ppm. The spectra of both the unadulterated and adulterated sample (Figures 3.a and 3.c, respectively) have in addition, some signals between 4 and 5.5 ppm which are not present in the spectrum of the pure dye. The common selected variables obtained with the three different methods (Figures 3.d-3.f) are placed in the aromatic zone between 6.9 and 9 ppm, which is the spectral zone where the Sudan II dye as well as Sudan I, III and IV dyes (not shown) present most of their signals. It must be emphasized that although the spectral region is the same, the variables are not exactly the same. Xdiff (Figure 3.d) looks

for individual variables that may or may not be consecutive; iPLS (Figure 3.e) looks for intervals of variables, such that there are consecutives within each interval, and finally GA (Figure 3.f) looks for consecutive variables within a short interval of variables, those resulting from expanding the selected variables in the last iteration. Evidence of this is a signal located at 8.086 ppm (behind the most intense peak found in this zone) which is only selected by GA. Moreover, GA and Xdiff also select variables located at the aliphatic zone.

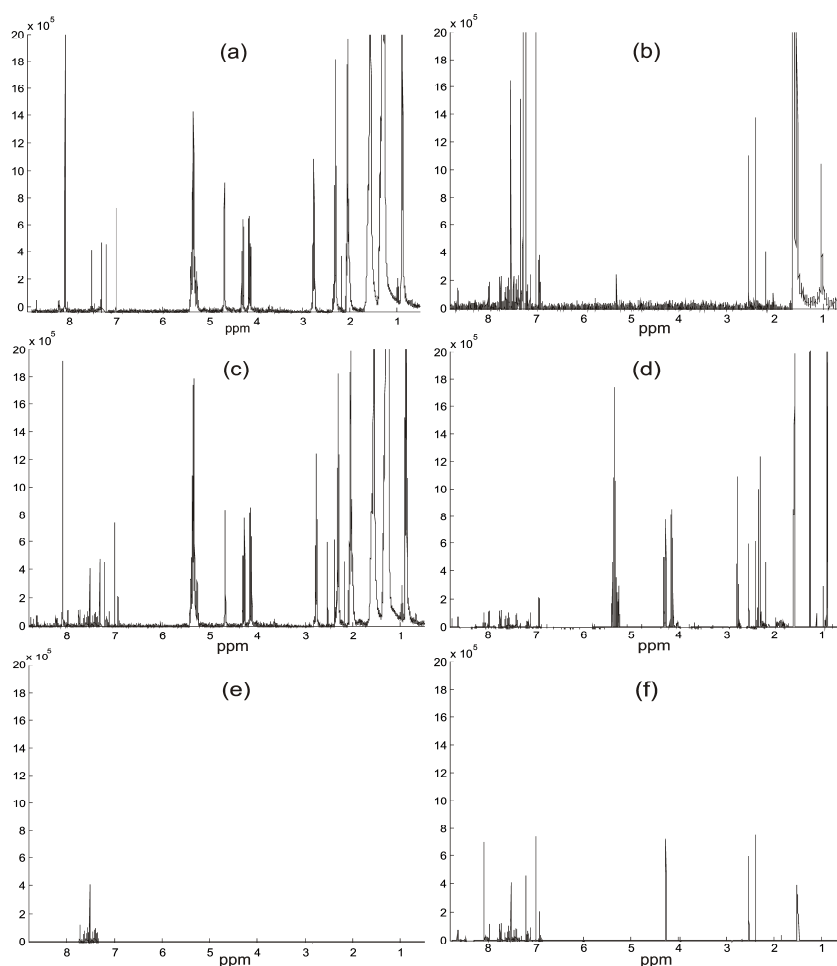


Figure 3. Spectrum of the (a) original variables of an unadulterated random sample, (b) the pure Sudan II dye, (c) the original variables of the sample adulterated with Sudan II and (d) variables selected with Xdiff, (e) iPLS and (f) GA methods.

Table 1 gives additional information which resumes the above discussion considering the first eight selected variables. It can be seen that the aromatic zone is the zone where most variables are selected by the three methods. Continuing with Table 1, the last three variables can be related to the NMR signals of the methyl groups which are only present in the chemical structure of the Sudan I and IV dyes [26]. This information related to the methyl groups is only considered by the first eight relevant variables selected by Xdiff although as it can be seen in figure 3.f, GA also selects some variables in the same zone that have lower relevance in the PLS loading weights.

Table 1: Proton chemical shifts for the first variables selected with Xdiff, iPLS and GA. The correspondence with the Sudan dyes signal is also indicated.

Chemical shift (ppm)	Sudan dyes presence	Xdiff	iPLS	GA
8.086	III			x
7.970	I-II-III-IV	x		
7.992	I-II-III-IV	x		
7.755	I-II-III-IV		x	
7.738	I-II-III-IV		x	x
7.715	I-II-III-IV		x	
7.596	I-II-III-IV		x	
7.563	I-II-III-IV	x		
7.519	I-II-III-IV		x	x
7.505	I-II-III-IV		x	
7.484	I-II-III-IV		x	
7.466	I-II-III-IV		x	
7.209	I-II-III-IV			x
6.995	I-II-III-IV			x
6.938	I-II-III-IV	x		
6.915	I-II-III-IV	x		
6.885	I-II-III-IV			x
5.989	----			x
4.270	----			x
2.534	II-IV	x		
2.385	II-IV	x		
2.380	II-IV	x		

Concerning the classification results, Table 2 shows the misclassified samples obtained by the three PLS-DA models built using the variables selected from each technique (Xdiff, iPLS and GA). The errors are discussed considering samples wrongly assigned (depicted in bold), samples not assigned to any class (called none) and samples assigned to more than one class (showing the two classes where these samples are assigned to). For the sake of clarity, when the above mentioned samples are correctly assigned with some of the models, the results are not shown (empty spaces).

Table 2. PLS-DA class assignment errors considering the variables selected by three techniques (Xdiff, iPLS and GA) for the training and test set. Wrongly assigned samples are depicted in bold, samples not assigned to any class are called as none and samples assigned to more than one class are presented (showing the two classes where these samples are assigned to). The sample number and its true class are indicated in the first two columns.

		Class assignment selecting variables with:				
		Sample	True class	<i>Xdiff</i>	<i>iPLS</i>	<i>GA</i>
TRAINING SET		1	1	1,5		
		2	1			1,5
		4	1			1,5
		6	1			1,5
		36	2	2,5		
		46	2		2,5	
		47	3	2	2	2
		80	4	4,5		
		81	4	4,5		
		82	4	4,5		
		96	5	None		None
		98	5		1	1
		100	5	None		
		101	5	5,4		
	108	5	5,4			
	113	5	5,4			
TEST SET		128	4		4,5	
		132	5	5,1		
		133	5	None		

Focusing at the training set, the first point concerns sample 47 which is the only wrong assigned sample using the three variable selection approaches. This sample belongs to class 3 and the percentage of predicted probability is 100% to class 2. So, in our opinion, this misclassification might be due to an experimental error and it can be considered as an outlier. A particular case is sample 98 that belongs to class 5 and it is wrongly assigned to class 1 with almost 99% of probability when using either GA or iPLS. However, it is correctly assigned only when using Xdiff. In our opinion this fact might be an indication that, in this case, some of the variables only selected by the Xdiff technique are crucial in obtaining the right classification.

There are some samples which are not assigned to any class (called none) either with Xdiff and/or GA. In some ways we think that it is better not to assign a sample to any class than to make a wrong assignation, particularly when dealing with foodstuff adulteration problems. It has to be stated that these samples are correctly assigned when using the iPLS technique.

The rest of samples shown in Table 2 correspond to samples assigned to more than one class, which have high predicted probability to belong to both of them and being one of the two classes its true class. Most of these samples are given when Xdiff technique is used, mainly for samples between class 4 (spices spiked with Sudan III) and class 5 (spices spiked with Sudan IV). On the other hand, the assignation results for the test set (Table 2) are similar to those presented above regarding the training set.

If the type of error and its implications are considered, some comments can be made: there are some unadulterated samples with also high probability to be assigned as contaminated with Sudan dyes, which implies that these samples must be discarded until their real status is confirmed. But the most important fact is

when the consumer health is under risk, i.e., samples contaminated with Sudan dyes which are classified as unadulterated ones. In our particular case, only one sample was obtained under this condition (sample 98). Finally, the general trend regarding the samples assigned to more than one class, is that there are adulterated samples with also high probability to belong to another adulterated class (classes 4 and 5), which has neither economical nor healthy implications as in any case they will be withdrawn from the markets.

The final overall recognition abilities regarding the training set are 87.8%, 95.7% and 99.1% for Xdiff, GA and iPLS, respectively. Considering the prediction abilities in the test set, a 90%, 100% and 95% is achieved by Xdiff, GA and iPLS, respectively. These results show that the use of an appropriate variable selection technique really improves the performance classification when dealing with NMR data.

5. Conclusions

We demonstrate that in a classification problem, when dealing with hundreds or thousands of variables as is the case of NMR data, the application of a variable selection approach is almost mandatory. The selection of variables, not only makes a considerable data reduction, but also eliminates noisy areas from the spectra or areas that do not contain relevant information in a way to achieve an optimal classification performance.

Although a visual inspection of the spectra in some cases, might allow selecting characteristics regions, it has been demonstrated that the application of variable selection techniques improves the classification results. This study focuses

on three different approaches and the following conclusions can be mentioned each one:

Xdiff is easy to implement, and has a clear and simple structure which allows individual variables to be selected. In this particular case, it is the only selection technique that selects variables across the entire spectra. An obvious example of this can be seen in the classification of sample 98, as it is only assigned correctly by this technique.

On the other hand, iPLS is conceptually and mathematically easy to implement and it is very effective at selecting the interesting parts of the spectrum. However, as it selects zones from across the entire spectra, it might also incorporate noisy variables, in such a way it cannot select punctual variables.

Finally, GAs may also be considered a good variable selection technique as this has been demonstrated. Several GA-based features selection methods using different GA structures have been developed, making this technique a flexible way to be applied in a variety of analytical problems. Nevertheless, it is conceptually more complex than the other two techniques studied and a larger number of parameters have to be considered as inputs to the algorithm.

As a final conclusion, based on the quite high correct classification results, the three variable selection techniques are very powerful tools in the determination of either food adulteration as well as the type of adulterant used when dealing with NMR data.

In our particular case study, iPLS and GA give better classification and prediction results. It has to be emphasized that most of the predictions errors are between two adulterated classes and there is only one case of an adulterated sample

being assigned unadulterated by iPLS and GA but not with Xdiff, which is of great importance whenever dealing with a food safety problem.

Acknowledgments

The authors would like to thank the Management Agency for University and Investigation Support of the Catalan Government (AGAUR) for providing Carolina Di Anibal a doctoral fellowship and the Spanish Ministry of Education, Culture and Sports (Project CTQ2007- 311 61474/BQU) for economic support.

6. References

- [1] IARC: International Agency for Research on Cancer, IARC monographs on the evaluation of the carcinogenic risk of chemicals to man: some aromatic azo compounds, Lyon, 1975, vol. 8, pp. 224–231.
- [2] C.V. Di Anibal, I. Ruisanchez, M.P. Callao, *Food Chem.* **124** (2011) 1139.
- [3] R.A. Davis, A.J. Charlton, S. Oehlschlager, J.C. Wilson, *Chemometr. Intell. Lab. Syst.* **81** (2006) 50.
- [4] J.W.E. Vogels, L. Terwel, A.C. Tas, F. van den Berg, F. Dukel, J. van der Greef, *J. Agric. Food Chem.* **44** (1996) 175.
- [5] D.W. Lachenmeier, E. Humpfer, F. Fang, B. Schütz, P. Dvortsak, C. Sproll, M. Spraul, *J. Agric. Food Chem.* **57** (2009) 7194.
- [6] R. Consonni, L.R. Cagliani, M. Stocchero, S. Porretta, *J. Agric. Food Chem.* **57** (2009) 4506.
- [7] B. Biais, J.W. Allwood, C. Deborde, Y. Xu, M. Maucourt, B. Beauvoit, W.B. Dunn, D. Jacob, R. Goodacre, D. Rolin, A. Moing, *Anal. Chem.* **81** (2009) 2884.
- [8] S. Rezzi, I. Giani, K. Héberger, D.E. Axelson, V.M. Moretti, F. Reniero, C. Guillou, *J. Agric. Food Chem.* **55** (2007) 9963.
- [9] S. Rezzi, D.E. Axelson, K. Héberger, F. Reniero, C. Mariani, C. Guillou, *Anal. Chim. Acta* **552** (2005) 13.
- [10] M. Cuny, E. Vigneau, G. Le Gall, I. Colquhoun, M. Lees, D.N. Rutledge, *Anal. Bioanal. Chem.* **390** (2008) 419.
- [11] G.R. Lloyd, K. Wongravee, C.J.L. Silwood, M. Grootveld, R.G. Brereton, *Chemometr. Intell. Lab. Syst.* **98** (2009) 149.
- [12] A. Jankevics, E. Liepinsh, E. Liepinsh, R. Vilskersts, S. Grinberga, O. Pugovics, M. Dambrova, *Chemometr. Intell. Lab. Syst.* **97** (2009) 11.
- [13] S.B. Kim, Z. Wang, S. Oraintara, C. Temiyasathit, Y. Wongsawat, *Chemometr. Intell. Lab. Syst.* **90** (2008) 161.
- [14] A.J. Charlton, P. Robb, J.A. Donarski, J. Godward, *Anal. Chim. Acta* **618** (2008) 196.
- [15] H. Winning, E. Roldán-Marín, L.O. Dragsted, N. Viereck, M. Poulsen, C. Sánchez-Moreno, M.P. Cano, S.B. Engelsen, *Analyst* **134** (2009) 2344.
- [16] H. Winning, N. Viereck, L. Nørgaard, J. Larsen, S.B. Engelsen, *Food Hydrocolloids* **21** (2007) 256.

-
- [17] H.W. Cho, S.B. Kim, M.K. Jeong, Y. Park, T.R. Ziegler, D.P. Jones, *Expert Syst. Appl.* **35** (2008) 967.
- [18] M. Wasim, R.G. Brereton, *Chemometr. Intell. Lab. Syst.* **81** (2006) 209.
- [19] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, *Appl. Spectrosc.* **54** (2000) 413.
- [20] D.B. Hibbert, *Chemometr. Intell. Lab. Syst.* **19** (1993) 277.
- [21] R. Leardi, *J. Chemometr.* **15** (2001) 559.
- [22] R. Leardi, A. Gonzalez Lupiáñez, *Chemometr. Intell. Lab. Syst.* **41** (1998) 195.
- [23] M. Barker, W. Rayens, *J. Chemometr.* **17** (2003) 166.
- [24] M. Bylesjö, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J. Trygg, *J. Chemometr.* **20** (2006) 341.
- [25] R.G. Brereton, *Trends Anal. Chem.* **25** (2006) 1103.
- [26] C.V. Di Anibal, M. Odena, I. Ruisanchez, M.P. Callao, *Talanta* **79** (2009) 887.
- [27] ASTA (American Spice Trade Association).
<http://www.astaspice.org/files/public/SudanWhitePaper.pdf>
- [28] <http://www.models.kvl.dk/source/>.
- [29] R. Leardi, *J. Chemometr.* **14** (2000) 643.

3.3.2. PAPER

¹H-NMR and UV-Visible Data Fusion for Determining Sudan Dyes
in Culinary Spices

Carolina V. Di Anibal, M. Pilar Callao, Itziar Ruisánchez

Talanta 84 (2011) 829-833

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

$^1\text{H-NMR}$ and UV-Visible Data Fusion for Determining Sudan Dyes in Culinary Spices

Carolina V. Di Anibal, M. Pilar Callao, Itziar Ruisánchez

Chemometrics, Qualimetrics and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n Tarragona, Spain.

Abstract

Two data fusion strategies (variable and decision level) combined with a multivariate classification approach (Partial Least Squares-Discriminant Analysis, PLS-DA) have been applied to get benefits from the synergistic effect of the information obtained from two spectroscopic techniques: UV-Visible and $^1\text{H-NMR}$. Variable level data fusion consists of merging the spectra obtained from each spectroscopic technique in what is called “meta-spectrum” and then applying the classification technique. Decision level data fusion combines the results of individually applying the classification technique in each spectroscopic technique. Among the possible ways of combinations, we have used the fuzzy aggregation connective operators. This procedure has been applied to determine banned dyes (Sudan III and IV) in culinary spices. The results show that data fusion is an effective strategy since the classification results are better than the individual ones: between 80 and 100% for the individual techniques and between 97 and 100% with the two fusion strategies.

Keywords: Variable level data fusion, Decision level data fusion, UV-Visible, $^1\text{H-NMR}$, Fuzzy aggregation connectives, Food adulteration.

1. Introduction

Nowadays, food quality and safety characterization is still one of the key issues whenever dealing with foodstuffs and great effort has been devoted to the detection of hazardous additives. The application of spectroscopic techniques has become a usual tool in food analysis [1] and requires the use and development of chemometrics tools in order to display and interpret vast amounts of data. There are some approaches that couple or merge data from two or more analytical techniques to improve multivariate data interpretation. This procedure goes under different names: data fusion, analysis of coupled or linked data, multiset or multiblock data analysis and integrative data analysis [2], among others. In this manuscript, we shall use the term data fusion as it is the term most commonly used by data analysts.

Data fusion has been applied in a variety of fields: for example, the combination of electronic noses and spectroscopic techniques to authenticate olive oil [3] and white grape must [4], and to determine sensory attributes in red wines [5]. Spectroscopic techniques have also been fused to identify pigments in works of art [6, 7] and cultivars of extra virgin olive oils [8]. Most fused data comes from Infrared (NIR, FT-IR, MIR), UV-Visible, Raman, Fluorescence, and Mass Spectrometry. One of the emerging research areas in which data fusion is applied is known as “metabolomics” or “metabonomics”, the goal of which is to obtain information from such highly complex samples as biofluids, cells and tissues [9, 10].

Previous studies have shown that the UV-Visible [11] and High-Resolution ^1H Nuclear Magnetic Resonance ($^1\text{H-NMR}$) [12] spectroscopic techniques coupled with multivariate classification techniques are well suited for determining the possible adulteration of commercial spices with Sudan dyes (I to IV). Based on those results, the main objective of the present study is to evaluate the combination of both techniques to improve the classification results. In addition, as different concentrations of Sudan dyes can be found in adulterated spices [13], in this paper we will explore the classification ability when the samples have lower concentration levels than those studied in the papers mentioned above. Also, as most of the classification errors were obtained with samples adulterated with Sudan III and IV, the present study focuses on these two dyes.

The idea is to get benefits of the possible synergism that two techniques as UV-Visible and NMR could have each other; due to the fact that both are based in different fundamentals and give different analytical signals, which allows thinking that the information provided by each one could be complementary.

The data provided by the two spectroscopic techniques have been processed separately and jointly by two data fusion strategies: variable and decision level data fusion. So, the overall performance of the classification process is evaluated through the well known classification technique Partial Least Squares-Discriminant Analysis (PLS-DA) for each individual spectroscopic technique and the fusion process.

2. Materials and Methods

2.1. Samples

A total of 35 spices from different common markets were studied. For UV-Visible analysis, each spice was extracted with acetonitrile and the obtained extract was twice filtered. For NMR, samples were dissolved in deuterated chloroform and once filtered. Samples contaminated with Sudan III and IV were prepared by spiking the non-contaminated samples at three concentration levels: 1.4, 3.6 and 7.1 g/Kg. In the end, then, three classes were defined: class 1 contained the 35 non-adulterated spices, class 2 contained a total of 105 samples adulterated with Sudan III corresponding to the three concentration levels (35 samples each one) and class 3 contained 105 samples corresponding to the three concentration levels adulterated with Sudan IV. More details about the sample treatment and experimental section can be found in previous studies [11, 12].

2.2. Spectrometric techniques and dataset

^1H -NMR spectra were acquired at 600.13 MHz on a Bruker Avance III-600 spectrometer, equipped with an inverse TCI 5 mm cryoprobe. One dimensional pulse experiments were carried out using a 90° pulse sequence (zg). For each sample, eight scans of 9.6 kHz of spectral width were collected at 300 K into 64 k data points. A recycling delay time of 15 s was applied between scans to ensure fully relaxation. Exponential line broadening of 0.3 Hz was applied before Fourier transformation and the NMR spectra acquired were phased, baseline-corrected (5th order polynomial adjustment) and calibrated by setting the CDCl_3 peak at 7.27 ppm (TopSpin 2.1, Bruker Biospin, Rheinstetten, Germany). UV-Visible measurements were made by an Agilent 8453 UV-Visible

spectrophotometer (Agilent Technologies Inc., Palo Alto, CA, USA) equipped with a diode array detector (DAD) and ChemStation Software (ChemStation Rev. A. 08.03). Each sample was measured against solvent as a blank in a 1cm-pathlength quartz cell and with a spectral resolution of 1 nm.

The UV-Visible spectra were acquired between 260 and 600 nm and had a total of 341 variables. The spectral region for NMR is located between 0.5 and 8.9 ppm and a range corresponding to the solvent signal (centred at 7.26 ppm) was removed, so finally, there were a total of 5698 variables.

The dataset (245 samples) is divided into a training and test set. The test set was generated by leaving out a 14% of the samples from class 1 and from each of the three concentration levels included in classes 2 and 3. The selection criterion is based on the PCA scores plot distribution from both UV-Visible and NMR data. Finally, the training set consisted of 210 samples and the test set of 35.

3. Chemometric Tools

3.1. Software

All chemometrics treatment was made with Matlab 6.5 Software (The MathWorks, Natick, MA) and PLS_Toolbox 3.5 (Eigenvector Research Incorporated).

3.2. Partial Least Squares-Discriminant Analysis

PLS-DA is the classical PLS regression technique adapted to a supervised classification task. A regression model is calculated that relates the independent

Chapter 3: Experimental Part and Results

variables (e.g. spectra) to an integer “ y ” that designates the class of the sample, with a binary response encoded as $\{1,0,0\}$ meaning that a sample belongs to class 1; $\{0,1,0\}$ to class 2 and $\{0,0,1\}$ to class 3. The model predicts the class for each sample based on a value from zero to one. A value close to zero indicates that the sample is not in the modelled class, while a value closer to one indicates that it is. A threshold between zero and one (above which a sample is considered part of the class) is calculated using Bayesian statistics [14]. The Bayesian threshold assumes that the “ y ” PLS predicted values are normally distributed and the threshold is selected at the y value at which the number of false positives and false negatives is minimized. More details of the PLS-DA technique can be found in the literature [15, 16].

The optimal number of latent variables (LVs) to be included in each model was chosen using leave-one-out cross-validation to minimize the root mean square-cross validation error (RMSECV) for each class. At the end, this number is selected through a compromise between the optimal value for each class.

3.3. Data fusion

Two levels of data fusion architectures are investigated in this paper: variable and decision level data fusion.

3.3.1. Variable level data fusion

Variable level fusion concatenates the variables into a single vector, which is called a “meta-spectrum”. Data must be balanced (all variables in the same scale) prior to the fusion process, so in our particular case only the NMR variables are

normalized since the UV-Visible intensity values are already between 0 and 1. If the number of concatenated variables is quite high, a variable selection is required. Of the various selection approaches, interval Partial Least Squares (iPLS) was used here [17]. We are not going to describe the iPLS methodology in detail, merely point out that it investigates the influential zones of the spectra that contain the most discriminating predictors, and calculates local PLS-DA models in pre-fixed narrow intervals.

3.3.2. Decision level data fusion

Decision level data fusion combines the classification results obtained from each individual technique. In this study, the PLS-DA classification results are fused using the fuzzy set theory which implements fuzzy aggregation connective operators. Fuzzy theory, introduced by Zadeh [18], is a powerful and general technology for processing information.

A fuzzy set allows membership values between 0 and 1, so in our case the PLS-DA class assignment values are normalized to the interval [0,1] through a simple rescaling such as the following equation [19]:

$$m_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})}$$

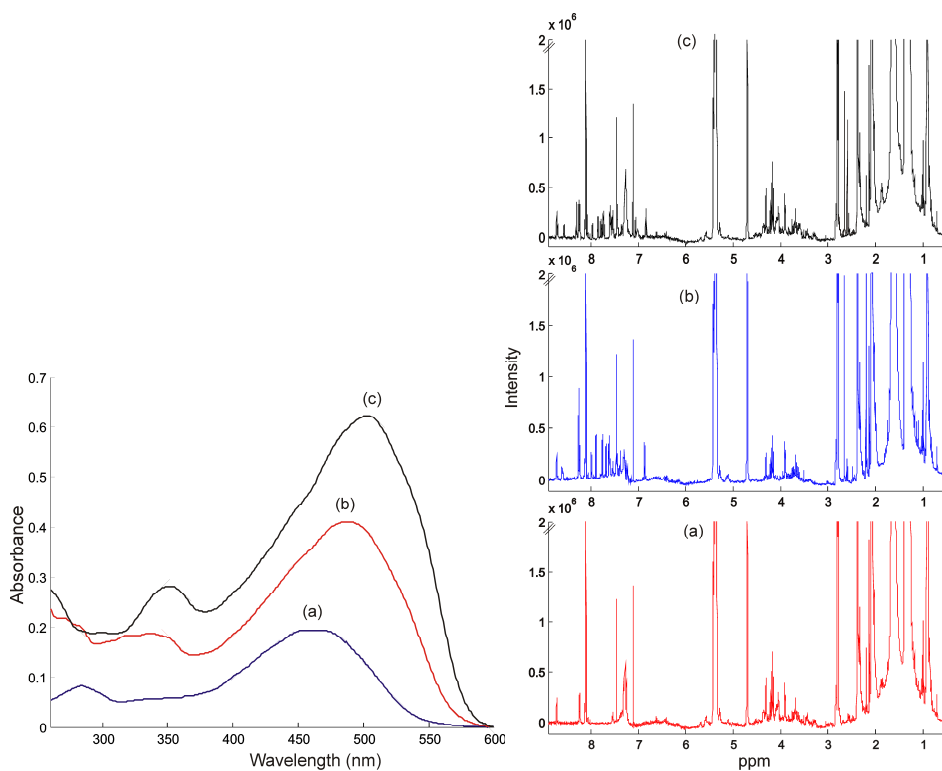
where $(m_{i,j})$ is the normalized class assignment value, $x_{i,j}$ is the PLS-DA class assignment value for the i th sample in the j th class and min and max are the minimum and maximum class assignment values in each j th class considering all samples, respectively.

Of the wide range of fuzzy connectives and aggregation operators that are available, we chose four aggregation connective operators which belong to the class of context independent constant behavior (CICB) operators [20]. The aggregation operators are: Minimum, Maximum, Product and Arithmetic Average. Considering the PLS-DA normalized assignment values obtained with each technique, the minimum and maximum are identified and the product and average are calculated. For the sake of clarity, two examples will be shown in the results section. To obtain the “ensemble decision” of each operator, the maximum value of the three possible classes is chosen [21]. The sample is finally assigned by the majority vote provided by all the fuzzy operators in the “ensemble decision”.

4. Results and Discussion

4.1. Spectra characterization

The only difference that exists between the chemical structures of the two Sudan dyes is the presence of two methyl groups that Sudan IV has. Figures 1 and 2 show the UV-Visible and NMR spectra, respectively, of (a) a random unadulterated paprika, (b) the same spice spiked with Sudan III and (c) the same spice spiked with Sudan IV. The UV-Visible spectra show that the absorption maximum from both adulterated samples shift slightly towards a longer wavelength respect to the unadulterated one, and that the sample adulterated with Sudan IV dye has the highest sensitivity (higher absorbance values). In the NMR spectra, a detailed comparison is not so easy, although it is evident that the aromatic zone in the samples containing Sudan dyes has more signals than the unadulterated sample.



Figures 1 (left) and 2 (right). UV-Visible and ¹H-NMR spectra, respectively, of a random paprika spice: (a) unadulterated, (b) spiked with Sudan III and (c) spiked with Sudan IV. Both Sudan dyes are at 7.1 g kg^{-1} (5 mg l^{-1} for UV-Visible and 50 mg l^{-1} for NMR).

4.2. Selection of test samples

Figure 3 shows, as an example for class 1, the PCA scores plots used for selecting the test set samples. In this case, 5 out of 35 samples are selected from both UV-Visible and NMR scores plots, to cover in the most representative way the PCA sample's distribution. For the other two classes, the same criterion has been applied but considering the concentration level, so 15 out of the 105 samples

from class 2 and 15 from out of the 105 samples from class 3 have been selected (5 from each concentration level).

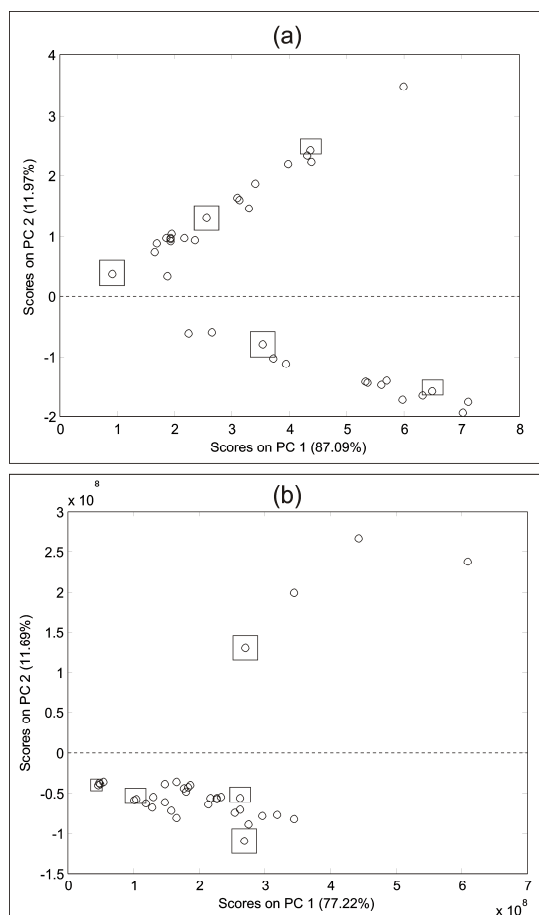


Figure 3. PCA scores plots from class 1: (a) UV-Visible and (b) NMR. Selected samples are marked with squares.

4.3. Independent decision-making

First of all, the UV-Visible and NMR data are pre-treated separately: the UV-visible spectra are mean centred while the NMR spectra are autoscaled. In

addition, before PLS-DA is applied as the classification technique, a NMR variable selection is carried out by means of the iPLS algorithm. The intervals selected by iPLS are depicted as solid-line rectangles (Figure 4) and it can be seen that these intervals are in both the aromatic zone and several aliphatic zones. The final number of selected variables is 777.

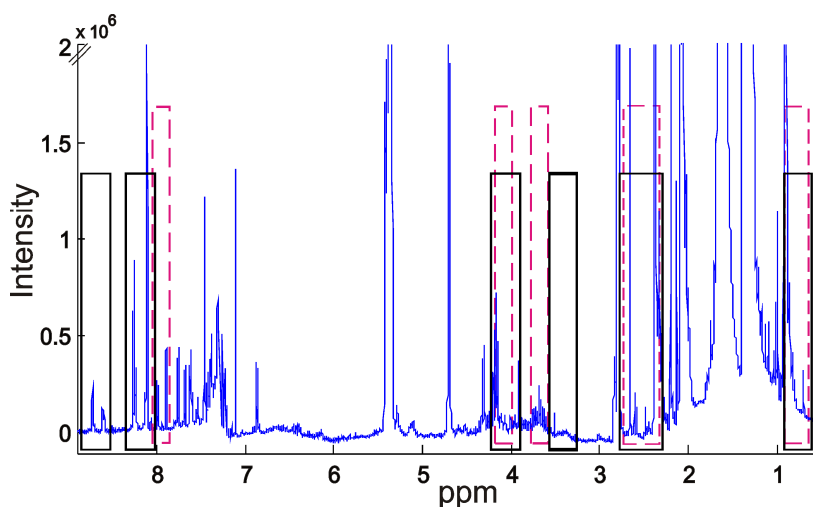


Figure 4. iPLS selected intervals for the NMR spectrum from a random spice spiked with Sudan III. The solid line represents the iPLS selected intervals when NMR is used individually. The dotted line represents the iPLS selected intervals when UV-Visible and NMR are fused.

Table 1 presents the PLS-DA misclassification results obtained from UV-Visible and NMR data independently. It can be seen that the overall error trend for both techniques is that samples are assigned to more than one class, one of which is the true class. It can also be seen that the misclassifications provided by the two techniques do not match each other, thus demonstrating the complementarities

existing between the different types of information. A closer look at the table shows that for the training set, most of the UV-Visible errors occur for samples from the unadulterated class (class 1), which also have a high probability (evaluated from the Bayesian distribution) of being assigned to class 3 (samples adulterated with Sudan IV). This probability varies between 99 and 100% for class 1 and between 79 and 100% for class 3. This type of error represents an economic risk since they must be withdrawn from commercial markets until they have been confirmed. Similarly, most of the NMR errors occur in samples from class 2 (adulterated with Sudan III), which are also assigned to the unadulterated class (class 1). To a lesser extent, the opposite also occurs: samples from class 1 are also assigned to class 2. The implications of these errors are the same as the ones discussed above, as they are samples assigned to more than one class.

Following the discussion of the table, only one sample is not assigned to any class with UV-Visible (sample 39) and two with NMR (samples 29 and 57), shown in hyphenated line (see Table 1). In addition, two samples are wrongly assigned to another class by NMR (9 and 33). Of these, sample 33 is an example of the most serious error possible, as it has implications for the consumer's health: a spice is being considered safe for human consumption when in fact it is not. Finally, if the proportion of samples in each data set is taken into account, the prediction results for the test set behave in a similar fashion to those of the training set. It should be mentioned that all the misclassified samples from classes 2 and 3 (Table 1) are adulterated at the lowest Sudan dye concentration.

Table 1. PLS-DA misclassification results when UV-visible or NMR are used individually.

	Sample	True class	PLS-DA class assignation	
			UV-visible	NMR
TRAINING SET	1	1	1, 3	
	3	1	1, 3	
	4	1	1, 3	
	5	1	1, 3	
	6	1	1, 3	
	7	1	1, 3	
	9	1		2
	11	1		1,2
	14	1	1, 2	
	18	1		1,2
	29	1		----
	30	1		1,2
	32	2		2,1
	33	2		1
	36	2		2,1
	38	2		2,1
	39	2	----	
	40	2		2,1
	50	2		2,1
	55	2		2,1
57	2		----	
127	3	2, 3		
TEST SET	211	1		1,2
	217	2		2,1
	219	2		2,1

4.4. Data fusion

4.4.1. Variable level fusion

Secondly, UV-Visible and NMR data are fused. For the variable level data fusion, the 341 and 5698 raw variables from UV-Visible and NMR, respectively, are concatenated into a meta-spectrum. Due to the fact that the number of variables is really high, three consecutive iPLS models were made. The final intervals selected include all the UV-Visible variables as well as some NMR zones, which are

depicted in Figure 4 as dotted line rectangles. At this point, 826 final variables are selected, 341 of which are UV-Visible and 485 NMR. The fact that the variables are selected from both spectroscopic techniques is an indication of the synergy between them: they both provide information that is important for discriminating the classes considered.

4.4.2. Decision level fusion

In the decision level data fusion, the individual class assignment results provided by PLS-DA are fused by means of the four fuzzy aggregation connective operators (minimum, maximum, product and average) and the majority vote rule, as described in the theory section. As an example, Table 2 shows the classification results in terms of numerical values when the different operators are used. The maximum values for each fuzzy operator are shown in bold. Sample 18 (class 1) is indistinctly assigned to classes 1 and 2 by NMR, but after the decisions of all the fuzzy operators it is correctly assigned. Similarly, sample 32 (class 2) is also indistinctly assigned to classes 1 and 2 by UV-Visible and NMR spectroscopic techniques, as they give scaled classification results very close each other. However, in this case the final sample classification is the same as when the individual techniques are used.

4.5. Comparison of the different strategies

Table 3 shows the PLS-DA misclassification results for the variable and decision level data fusion. Overall, it can be seen that there is a great improvement in the results provided by the two fusion strategies, since the number of errors is

lower than when the individual techniques are used (Table 1). With variable level data fusion no wrong assignments to another class are obtained. These results are positive because, from a practical point of view, samples that are assigned to more than one class or assigned to any class should be confirmed by an alternative technique. This is preferable to assigning a sample wrongly. With decision level data fusion, two adulterated samples (33 and 57) pose a serious problem because, as mentioned above, they can affect the consumer's health. The other misclassified sample (32) become in a suspicious sample that has to be subsequently confirmed as mentioned above. As in the previous case, the test set follows the same pattern as the training set.

Table 2. Assignment of classes by decision level data fusion.

Sample 18	Scaled PLS-DA class assignment values			Ensemble decisión
	Class 1	Class 2	Class 3	
UV-Visible	0.68	0.42	0.23	
NMR	0.51	0.51	0.30	
Minimum	0.51	0.42	0.23	Class 1
Maximum	0.68	0.51	0.30	Class 1
Product	0.34	0.21	0.07	Class 1
Average	0.59	0.46	0.26	Class 1
Majority Vote				Class 1

Sample 32	Scaled PLS-DA class assignment values			Ensemble decisión
	Class 1	Class 2	Class 3	
UV-Visible	0.46	0.45	0.19	
NMR	0.48	0.50	0.33	
Minimum	0.46	0.45	0.19	Class 2,1
Maximum	0.48	0.50	0.33	Class 2,1
Product	0.22	0.23	0.06	Class 2,1
Average	0.47	0.48	0.26	Class 2,1
Majority Vote				Classes 2,1

Table 3. PLS-DA misclassification results for variable and decision level data fusion.

	Sample	True class	PLS-DA class assignation	
			Variable fusion	Decision fusion
TRAINING SET	3	1	1, 3	
	32	2		2,1
	33	2	-----	1
	57	2		1
	127	3	-----	
TEST SET	235	3	-----	

Finally, Table 4 shows the correct classification percentages obtained with the individual spectroscopic techniques and with the two fusion strategies. The global classification percentages obtained for class 1 with variable and decision level fusion strategies are clearly better than those obtained with UV-Visible and NMR. For class 2, the fusion strategies improve the classification results obtained with the NMR technique while for class 3 all four strategies give satisfactory and comparable classification values. A detailed look at the results considering the concentration levels (classes 2 and 3) indicates that lower percentages are obtained for samples at the lowest concentration level, being the worst case the percentage obtained for class 2 with the individual NMR technique (71,4%). For class 2, the NMR classification result is improved by the two fusion strategies which give a 97 and 91.4% of correct classification with variable and decision fusion respectively, which are similar to the UV-Visible one. For class 3, the percentages obtained with the individual techniques and fusion strategies are similar although the variable selection strategy gives lower values than the individual ones. On the other hand, with the other two concentration levels, the maximum classification ability is obtained (100%) in all cases (individual techniques and data fusion strategies).

Table 4. Global correct classification percentages for each class (bold values). For classes 2 and 3, the results corresponding to each concentration level (from lower to upper) are shown in brackets.

	Class 1	Class 2	Class 3
UV-Visible	80.0	99.0 (97.1; 100; 100)	99.0 (97.1; 100; 100)
NMR	82.8	90.5 (71.4; 100; 100)	100 (100; 100; 100)
Variable fusion	97.1	99.0 (97.1; 100; 100)	98.1 (94.3; 100; 100)
Decision fusion	100	97.1 (91.4; 100; 100)	100 (100; 100; 100)

5. Conclusions

Fusing data from UV-Visible and $^1\text{H-NMR}$ instruments is a powerful tool for detecting banned Sudan dyes at three different concentrations levels in commercial spices destined for human consumption. None of the adulterated samples with Sudan III and Sudan IV were misclassified to each other. Some samples at the lowest concentration level were assigned to their own class and also to the non adulterated class.

Tacking into account that nowadays many laboratories have a variety of analytical equipment any data fusion strategy is a feasible way to deal with multivariate approach. Decision level data fusion has the extra-advantage that it can be applied to all types of measurements, since it combines the individual multivariate results. Fuzzy aggregation connectives have been demonstrated to be a good and simple tool to implement for classification analysis.

The benefits of the data fusion methodology in the present study are clear because the classification results are better than the ones obtained individually with UV-Visible and NMR techniques, thus demonstrating that the information obtained from the two spectroscopic techniques has a synergistic effect.

Acknowledgments

The authors would like to thank the Agency for the Administration of University and Research Grants of the Catalan Government (AGAUR) for providing Carolina Di Anibal with a doctoral fellowship.

6. References

- [1] C. Cordella, I. Moussa, A.C. Martel, N. Sbirrazzuoli, L. Lizzani-Cuvelier, *J. Agric. Food Chem.* **50** (2002) 1751.
- [2] I. V. Mechelen, A. Smilde, *Chemometrics Intell. Lab. Syst.* **104** (2010) 83.
- [3] M. Casale, C. Casolino, P. Olivieri, M. Forina, *Food Chem.* **118** (2010) 163.
- [4] S. Roussel, V. Bellon-Maurel, J.M. Roger, P. Grenier, *J. Food Eng.* **60** (2003) 407.
- [5] D. Cozzolino, H.E. Smyth, K.A. Dattey, W. Cynkar, L. Janik, R.G. Damberg, I. Leigh Francis, M. Gishen, *Anal. Chim. Acta* **563** (2006) 319.
- [6] P. Ramos, I. Ruisanchez, K. Andrikopoulos, *Talanta* **75** (2008) 926.
- [7] P. Ramos, I. Ruisanchez, *Anal. Chim. Acta* **558** (2007) 274.
- [8] M. Casale, N. Sinelli, P. Oliveri, V. Di Egidio, S. Lanteri, *Talanta* **80** (2010) 1832.
- [9] J. Forshed, H. Idborg, S.P. Jacobsson, *Chemometrics Intell. Lab. Syst.* **85** (2007) 102.
- [10] A.K. Smilde, M.J. van der Werf, S. Bijlsma, B.J.C. van der Werff-van der Vat, R.H. Jellema, *Anal. Chem.* **77** (2005) 6729.
- [11] C.V. Di Anibal, M. Odena, I. Ruisanchez, M.P. Callao, *Talanta* **79** (2009) 887.
- [12] C.V. Di Anibal, I. Ruisanchez, M.P. Callao, *Food Chem.* **124** (2011) 1139.
- [13] K. Mishra, S. Dixit, S.K. Purshottam, R.C. Pandey, M. Das, S.K. Khanna, *Int. J. Food Sci. Technol.* **42** (2007) 1363.
- [14] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, USA (2006).
- [15] M. Barker, W. Rayens, *J. Chemometr.* **17** (2003) 166.
- [16] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Kich, PLS_Toolbox Version 3.5 for use with Matlab™, Eigenvector Research, Inc., Manson, WA, USA (2005) p. 185-189.
- [17] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, *Appl. Spectrosc.* **54** (2000) 413.
- [18] L.A. Zadeh, *Inf. Control* **12** (1968) 94.
- [19] P. Ramos, M.P. Callao, I. Ruisanchez *Anal. Chim. Acta* **584** (2007) 360.
- [20] I. Bloch, *IEEE Trans. Syst. Man Cybern. Part A* **26** (1996) 52.
- [21] L. Lam, S.Y. Suen, *IEEE Trans. Syst. Man Cybern. Part A* **27** (1997) 553.

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

3.3.3. PAPER

Standardization of UV-Visible data in a food adulteration
classification problem

Carolina V. Di Anibal, Itziar Ruisánchez, Mailén Fernández, Rafel Forteza, Victor
Cerdà, M. Pilar Callao

Submitted

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

3.3.3. Standardization of UV-Visible Data in a Food Adulteration Classification Problem

Carolina V. Di Anibal¹, Itziar Ruisánchez¹, Mailén Fernández², Rafel Forteza²,
Victor Cerdà² M. Pilar Callao¹

¹Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n Campus Sescelades, E-43007 Tarragona, Spain

²Department of Chemistry, Faculty of Sciences, University of the Balearic Islands, Carretera de Valldemossa, Km 7.5. E-07122 Palma de Mallorca, Spain.

Abstract

This study evaluates the performance of multivariate calibration transfer methods in a classification context. The spectral variation caused by some experimental conditions can worsen the performance of the initial multivariate classification model but this situation can be solved by implementing standardization methods such as Piecewise Direct Standardization (PDS). This study looks at the adulteration of culinary spices with banned dyes such as Sudan I, II, III and IV. The samples are characterized by their UV-Visible spectra and Partial Least Squares–Discriminant Analysis (PLS-DA) is used to discriminate between unadulterated samples and samples adulterated with any of the four Sudan dyes. Two different datasets that need to be standardized are presented. The

standardization process yields positive classification results comparable to those obtained from the initial PLS-DA model, in which high classification performance was achieved.

Keywords: Multivariate standardization, Transfer, PDS, PLS-DA, Food adulteration

1. Introduction

Multivariate classification methods aim to identify relationships among objects and establish classification rules for discrimination purposes. Several multivariate classification methods have been proposed based on different mathematical approaches and they have been used in many fields, such as food science [1], metabolomics [2] pharmaceuticals [3], biomedicine [4] and environmental sciences [5]. Particularly, the food field has focused on detecting adulteration or contamination, authenticating claims, determining geographical origin and differentiating and classifying varieties, to name just a few applications. A review focusing on this context can be found elsewhere in the literature [6].

When a multivariate model is established, it is characterized by its ability to predict future samples. Over time, the validity of the model must be tested to ensure that this ability is maintained, as there are some situations in which the model might not be valid. These situations include some changes in experimental conditions.

There are basically two solutions for dealing with the aforementioned problem. The first one consists of performing a new model, but this is not a practical solution because it implies an increase in both time and cost. Another

useful alternative is to make use of the chemometric methods known as standardization or calibration transfer methods. These methods try to correct the differences between the measurements performed under different conditions, being one of them the condition used to build the original multivariate model, in order to make this model useful at a lower experimental cost [7].

Several algorithms have been developed for this purpose, such as single wavelength standardization (SWS) [8,9], slope/bias correction (SBC) [10,11], direct and piecewise direct standardization (DS, PDS) [12-15], the Shenk-Westerhaus method [16,17], a two-block PLS approach [18], wavelet transform based-standardization [19,20], and artificial neural networks [21,22]. Pre-processing techniques have also been developed such as Orthogonal Signal Correction (OSC), Finite Impulse Response (FIR), Multiplicative Signal Correction [23], as well as modifications to some of the methods named above [24-28]. Many of these methods have been used mainly in a calibration context with infrared (IR) data.

Little research has been published regarding the application of standardization techniques in a classification context, although a few works can be cited [26, 28-32]. The scarce use of standardization techniques in classification contexts compared to its application in the calibration field might be due to the different types of results derived from each technique. The latter case yields quantitative results which provide better evidence of prediction errors. Meanwhile, in classification situations, the results are qualitative in nature (sample assignation to pre-defined groups) and these types of errors might not be so evident.

The objective of this paper is to evaluate the application of standardization methods in a food adulteration problem. The idea is to promote and show the potential of such methods in classification problems, as it is advantageous from a

practical point of view to maintain the robustness of the initial model in routine analysis.

This study addresses the adulteration of culinary spices with banned dyes such as Sudan I, II, III and IV. The samples were characterized by their UV-Visible spectra and Partial Least Squares–Discriminant Analysis (PLS-DA) was used to discriminate between unadulterated samples and samples adulterated with any of the four Sudan dyes [33]. PDS was chosen as standardization technique because it allows both intensity differences and wavelength shifts to be corrected [7]. The transfer samples were selected by means of the Kennard-Stone algorithm.

2. Theory

2.1. Selection of transfer samples

The Kennard-Stone algorithm [34] selects one by one the samples which are furthest from each other in the group, so they spread throughout the multivariate space they determine. The term used to measure the distance is the Euclidean distance. For a matrix with N samples and K variables, the Euclidean distance between samples i and j (D_{ij}) is defined as:

$$D_{ij} = \sqrt{\sum_{v=1}^k (x_{iv} - x_{jv})^2}$$

The two first selected samples are the two furthest samples (maximum D_{ij}). The third sample is selected according the following steps: the distance between each sample and the previous selected samples is calculated; the shortest of each of these distances is chosen and the sample with the maximum value in this group of

minimal distances is selected. The same criterion is applied for the rest of samples to be selected.

2.2. Piecewise Direct Standardization (PDS)

PDS [12] establish a mathematical relationship between the spectra of a sample measured under two different experimental conditions, in which the response of each i variable under the first condition r_{1i} is related to the response of a group of variables $\mathbf{R}_2 = [r_{2i-j}, \dots, r_{2i}, \dots, r_{2i+j}]$ under the second condition. Using all the transfer samples, the relationship can be written as:

$$\mathbf{r}_{1i} = \mathbf{R}_{2i}\mathbf{b}_i + \mathbf{b}_{oi}$$

where the coefficients \mathbf{b}_i and \mathbf{b}_{oi} are obtained from a multivariate technique such as PCR or PLS.

If the process is repeated for all wavelengths, n vectors of b coefficients and n values of b_o are obtained, which are grouped in the corresponding matrix and vector:

$$\mathbf{F} = \text{diag}(\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_i, \dots, \mathbf{b}_n)$$

$$\mathbf{b}_o^T = (b_{o1}, b_{o2}, \dots, b_{oi}, \dots, b_{on})$$

Both matrix and vector are used to correct a sample measured under the second condition (\mathbf{r}_2):

$$(\mathbf{r}_2)^{\text{std}} = \mathbf{r}_2^T \mathbf{F} + \mathbf{b}_o^T$$

where $\mathbf{r}_2^{\text{std}}$ is the standardized spectrum. Once the spectrum has been corrected, this sample can be classified into the initial multivariate model.

2.3. Partial Least Squares-Discriminant Analysis (PLS-DA)

Although it has been developed as a regression method, PLS can be applied to solve classification problems. This approach uses a linear multivariate model to relate independent variables (e.g. spectra) to dependent variables that designates the class of the sample by means of a binary code, in which 1 indicates that the sample belongs to the class of interest and 0 indicates that it belongs to a different class. A more detailed description of this technique is provided elsewhere [35].

3. Materials and Methods

3.1. Samples

The description of the three datasets is presented in Table 1. The data correspond to unadulterated culinary spices and spices adulterated with Sudan I to IV dyes. Class 1 contains the unadulterated samples while classes 2, 3, 4 and 5 contain the samples adulterated with Sudan I, II, III and IV, respectively. From one condition to another the same spectrometer is used; and the following parameters have been modified: (1) different operator and (2) the period of time in which the samples are analysed. The changes derived from such experimental parameters are not controlled and might have a different impact on the UV-Visible measurements.

Table 1. Description of datasets.

Dataset	Classes	Number of samples	Food matrix	Operator	Time (months)
1- First condition: Master data	1	135 (27 x 5)	Turmeric	1	0
	2		Curry		
	3		Hot paprika		
	4		Mild paprika		
	5				
2- Second condition	1	126 (42 x 3)	Turmeric	2	9
	4		Curry		
	5		Hot paprika		
			Mild paprika		
3- Third condition	1	60 (20 x 3)	Hot paprika	3	15
	2		Mild paprika		
	4				

3.2. Methods

UV-Visible spectra were recorded on an Agilent 8453 UV-Visible spectrophotometer (Agilent Technologies Inc., Palo Alto, CA, USA) equipped with a diode array detector (DAD) and ChemStation software. Each sample was measured against solvent as a blank in a 1cm-pathlength quartz cell and with a spectral resolution of 1 nm in the range between 260 and 600 nm. Reagents and experimental conditions have been reported elsewhere [33].

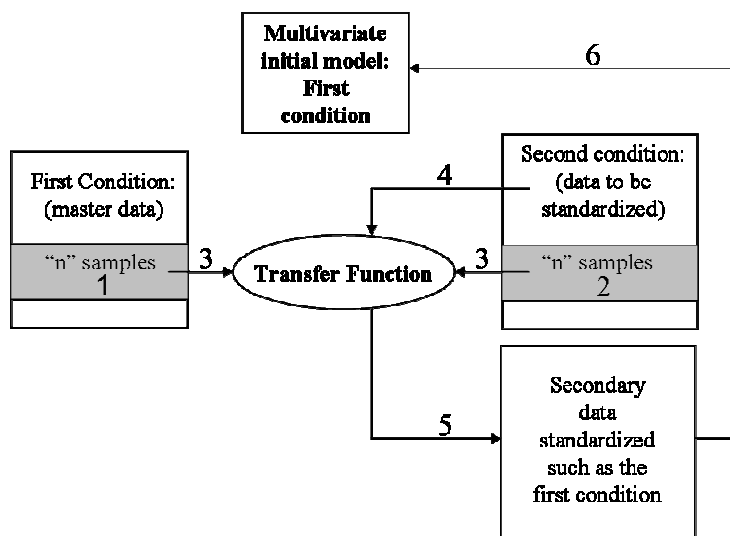
The PLS-DA classification model has been established in other work [33] and was performed using six latent variables that were chosen because of the minimum RMSECV (root mean square-cross validation error) for all classes. Data was mean-centred before the multivariate analysis.

PDS and the Kennard-Stone algorithm were performed using the Matlab PLS_Toolbox [36]. For PDS function, the local models were calculated with additive background correction and the default tolerance of the PLS Toolbox was employed (10^{-2}).

3.3. Standardization process

The standardization process is presented in Scheme 1. The multivariate model was built under a first condition and the samples to be predicted correspond to another condition (second condition). The first step (1) is to select the "n" transfer samples from the master data, and in a second step (2) these "n" samples are measured under the current conditions (second condition). Next, in a third step (3), a transfer function is built by using both groups of samples and in the fourth step (4) this function is used to transform the data obtained in the second condition (step 5). Finally, the transformed (standardized) data is ready to be predicted by applying the initial multivariate model (step 6).

Scheme 1. Steps of the standardization process.



4. Results and Discussion

For the sake of clarity considering the structure of the datasets, we will first discuss the standardization process applied to the dataset obtained under the third condition and then discuss the one obtained under the second condition. The optimization of the parameters for the standardization process has been evaluated independently for both datasets considering the global classification results. The number of transfer samples was varied from 6 to 15 (two to five from each class) and the samples were selected considering each class individually. Four PDS window sizes were tested: 3, 5, 7 and 9.

4.1. Standardization for third condition data

When the samples from this dataset were predicted by means of the classification model established under the first condition, they present a global classification error of about 50% compared to the initial data where no misclassifications were found.

As it has been previously stated, the “n” transfer samples are independently selected by means of the Kennard-Stone algorithm for each studied class. For example, to view the distribution of the “n” selected transfer samples, a PCA scores plot in the space defined by the two first principal components is shown in Figure 1 (the example correspond to class 1). The selected samples have been circled and the order in which they were selected is also indicated. It can be seen that the selection is representative as these transfer samples span the PCA scores-space.

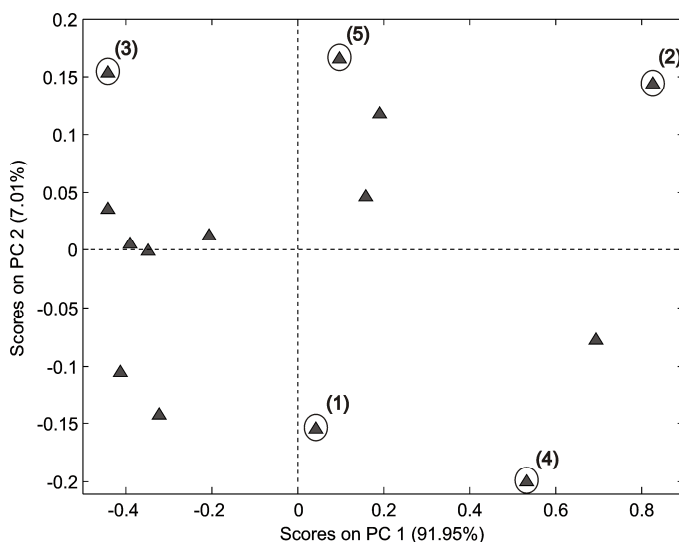


Figure 1. PCA scores plot distribution for the selection of the first five Kennard-Stone samples from class 1. The standardization of data under the second condition is considered. The selection order is also depicted.

When both the number of transfer samples and window sizes vary, the classification results remain the same. The possibility of using the minimum number of transfer samples represents a clear advantage, as fewer standardization samples mean less time and effort spent on analyses and, consequently, lower costs. Therefore, six samples (two from each class) with an intermediate window size value such as five were selected as the final parameters.

Figure 2 shows, as an example, the UV-Visible spectra for a random paprika sample from class 4 (sample spiked with Sudan III) obtained under the first and third conditions, along with the corresponding standardized spectrum. As the figure shows, the spectrum from the third condition has the same basic shape but it is shifted-down compared to the original spectrum (first condition). Otherwise, after the standardization it can be observed that the spectrum resembles the first condition.

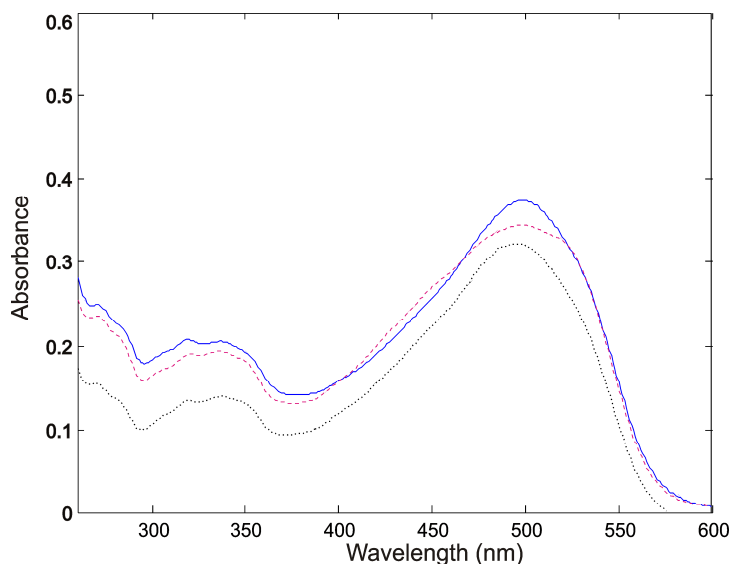


Figure 2. Spectra of a random paprika sample from class 4 obtained under the first condition (solid-line), the third condition (dotted-line) and after standardization (dashed-line).

In order to compare the classification results obtained before and after the standardization process, Figure 3 depicts a bar graph that summarizes this information along with the original classification results (master data) on each class studied. As the figure shows, after the standardization there is a great improvement in the classification results because a high performance is obtained for all of the classes. The only misclassification found correspond to an adulterated sample that is assigned to two classes, its own and another adulterated class.

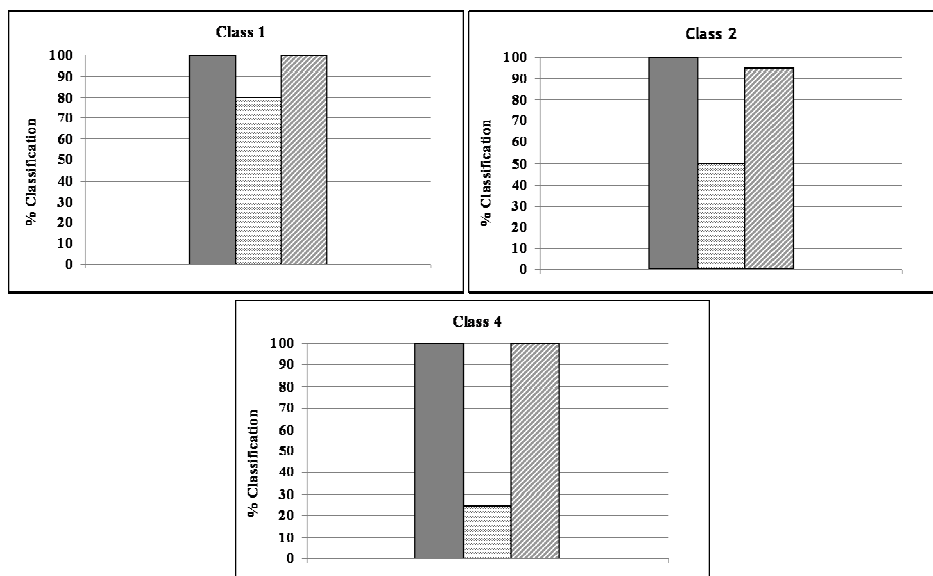


Figure 3. Classification results comparing data before and after standardization with data obtained under the third condition. The results from the master data are also represented. Full bars (master data), dotted bars (non standardized data) and dashed bars (standardized data).

4.2. Standardization for second condition data

The overall classification result for this dataset using the established model is 74%, so a transfer function was built following the same scheme presented in section 3.3. After the standardization, the global classification results were 17% lower compared to the results for the non-standardized data.

Looking at the spectrum of a paprika sample obtained under the second condition (Figure 4), it can be seen that, compared to the first condition, it is partially shifted up in the zone between approximately 400-500 nm. This spectral

zone is the least selective zone for pure Sudan III and IV dyes, as it can be appreciated from their UV-Visible spectra shown in Figure 5.

Otherwise, when looking at the spectrum of a curry sample (Figure 6), it can be observed that the entire spectrum is shifted up compared to the one obtained under the first condition (the same behaviour is observed for the turmeric samples). This situation suggests that depending on whether it is a paprika or a turmeric/curry sample, the spectral differences are not the same.

In light of the aforementioned differences found between the species, we decided to implement two transfer functions: one for each group of spices (curry/turmeric and mild/hot paprika). The “n” transfer samples were individually selected for each group of spices contained in each class.

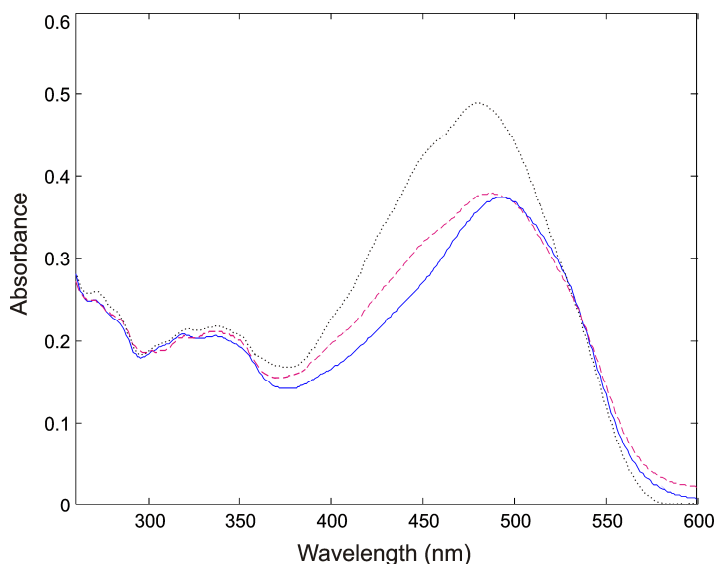


Figure 4. Spectra of the same paprika sample shown in Figure 2 obtained under the first condition (solid-line), the second condition (dotted-line) and after standardization (dashed-line).

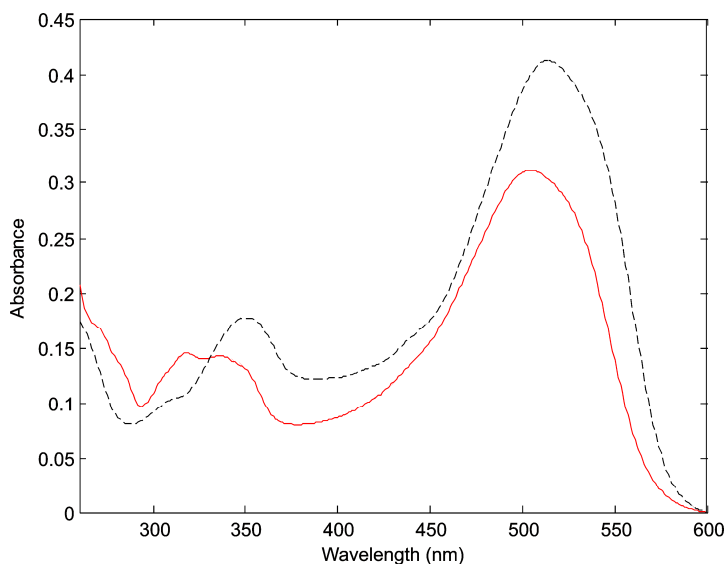


Figure 5. Spectra of pure Sudan III (solid-line) and Sudan IV (dashed-line) dyes.

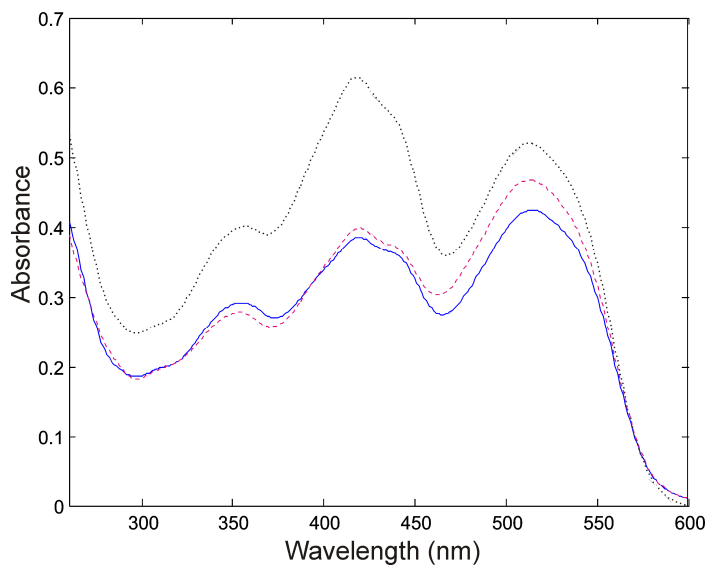


Figure 6. Spectra of a curry sample from class 5 obtained under the first condition (solid-line), the second condition (dotted-line) and after standardization (dashed-line).

For the sake of clarity, we will first discuss the mild/hot paprika group and secondly the curry/turmeric group. The same trend followed by the samples from the third condition emerges when evaluating the optimal standardization parameters: the classification results do not vary when different number of transfer samples and window sizes is used. Following the same criteria, the minimum number of transfer samples (six samples, two from each class) with a five window size were finally selected.

Figure 4 shows the UV-Visible spectra for a paprika sample obtained under the first and second condition along with the standardized spectrum. It can be observed that the transformed spectrum closely fits the profile of the first condition, which is more evident in the higher-absorbance zone of the spectrum.

Considering the classification results before and after the standardization, Figure 7 shows a bar graph for each class and group considered. The non-standardized data belonging to classes 1 and 4 are correctly classified with the initial classification model (first condition), so such data do not need to be standardized. After the standardization process, class 5 improve the classification results from 78.5 to 100% whereas the results for classes 1 and 4 remain the same. This suggests that although the standardization process is necessary for one of the three classes, if it is implemented for all three it does not negatively affect the classification results.

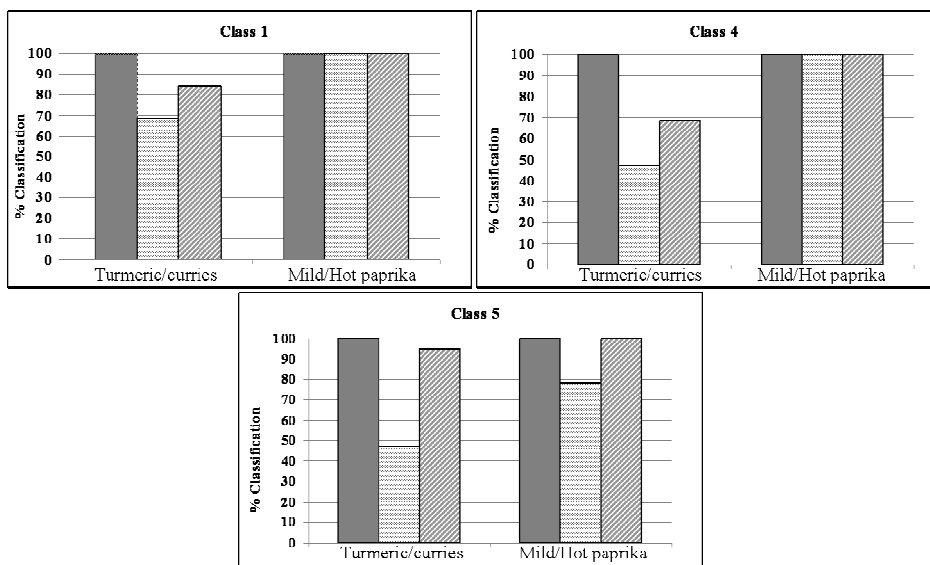


Figure 7. Classification results comparing data before and after standardization with data obtained under the second condition. The results from the master data are also represented. Full bars (master data), dotted bars (non standardized data) and dashed bars (standardized data).

Following the discussion for the other group of spices (turmeric/curries), the evaluation for the optimal standardization parameters is presented in Table 2. As a general trend, slight differences in the classification results were found, so the choice of an optimal condition that satisfies all the classes is impossible (see bold values). Therefore, the minimum number of transfer samples (six samples, two from each class) with a nine window size were chosen because the first parameter is the optimal for two out of three classes and the second is the best choice for all the classes.

Table 2. Optimization of the number of transfer samples and PDS window size on data obtained under the third condition (curries/turmeric group). All three classes are represented. The optimal standardization parameters for each class are shown in bold.

Class 1				
Number of transfer samples	% Classification error			
	win=3	win=5	win=7	win=9
2	73.68	78.95	78.95	84.21
3	73.68	73.68	73.68	78.95
4	73.68	73.68	78.95	73.68
5	78.95	84.21	84.21	84.21

Class 4				
Number of transfer samples	% Classification error			
	win=3	win=5	win=7	win=9
2	73.68	68.42	68.42	68.42
3	73.68	73.68	73.68	73.68
4	84.21	84.21	84.21	84.21
5	78.95	78.95	78.95	78.95

Class 5				
Number of transfer samples	% Classification error			
	win=3	win=5	win=7	win=9
2	94.74	94.74	94.74	94.74
3	78.94	73.68	73.68	73.68
4	78.95	78.95	78.95	78.94
5	89.47	84.21	84.21	84.21

Figure 6 shows a random curry sample from class 5 obtained under the first and second conditions and the corresponding standardized spectrum. The figure shows that, although there is a pronounced difference between the spectra belonging to the different conditions (shifted up), after the standardization it resembles the spectrum measured under the first condition.

Finally, regarding the classification results (bar graphs, Figure 7) the non-standardized data does not yield satisfactory classification performance (global

results of less than 50%), however these results improve after standardization, which mainly affects classes 4 and 5. Although such results do not reach the classification levels obtained in the first condition, positive results are derived from the standardization process, as there are no classification errors that would have an impact on consumer health (adulterated samples assigned as safe to be consumed) unlike when non-standardized data is evaluated.

5. Conclusions

Standardization methods have been proved to be valuable tools to preserve the performance of a multivariate classification model, as they can save both time and money.

PDS allowed the UV-Visible spectra recorded under different conditions to resemble the spectra registered under the initial condition in such a way that the classification model performs properly. As a general trend, the standardization process yields comparable results to those obtained from the initial PLS-DA model, whose results were quite satisfactory.

When dealing with a multiclass classification problem, sometimes not all the spectra from the different classes need to be standardized, nevertheless, it has been demonstrated that the standardization process does not worsen the classification results in such case. We can conclude that, overall, the standardization process is positive.

In the particular case of culinary spices adulterated with Sudan dyes, we found that depending on the type of spice studied (turmeric/curries or mild/hot

paprika), different transfer functions have to be considered when spectra have to be modified.

Acknowledgments

The authors would like to thank the Agency for the Administration of University and Research Grants of the Catalan Government (AGAUR) for providing Carolina Di Anibal with a doctoral fellowship.

6. References

- [1] C. Cordella, I. Moussa, A.C. Martel, N. Sbirrazzuoli, L. Lizzani-Cuvelier, *J. Agric. Food Chem.* **50** (2002) 1751.
- [2] R. Madsen, T. Lundstedt, J. Trygg, *Anal. Chim. Acta* **659** (2010) 23.
- [3] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, *J. Pharm. Biomed. Anal.* **44** (2007) 683.
- [4] M.R. Guarracino, S. Cuciniello, D. Feminiano, G. Toraldo, P.M. Pardalos, *CRM Proc. Lecture Notes* **45** (2008) 109.
- [5] S. Mas, A. de Juan, R. Tauler, A.C. Olivieri, G.M. Escandar, *Talanta* **80** (2010) 1052.
- [6] L.A. Berrueta, R.M. Alonso-Salces, K. Héberger, J. *Chromatogr. A* **1158** (2007) 196.
- [7] R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferré, *Chemom. Intell. Lab. Syst.* **64** (2002) 181.
- [8] F. Sales, M.P. Callao, F.X Rius, *Chemom. Intell. Lab. Syst.* **38** (1998) 63.
- [9] L. Norgaard, *Chemom. Intell. Lab. Syst.* **29** (1995) 283.
- [10] B.G. Osborne, T. Fearn, *J. Food Technol.* **18** (1983) 453.
- [11] E. Bouveresse, C. Hartmann, D.L. Massart, I.R. Last, K.A. Prebble, *Anal. Chem.* **68** (1996) 982.
- [12] Y.D. Wang, D.J. Veltkamp, B.R Kowalski, *Anal. Chem.* **63** (1991) 2750.
- [13] T.M. Alam, M.K. Alam, S.K. McIntyre, D.E. Volk, M. Neerathilingam, B.A. Luxon, *Anal. Chem.* **81** (2009) 4433.
- [14] E. Bouveresse, D.L. Massart, *Chemom. Intell. Lab. Syst.* **32** (1996) 201.
- [15] S. Macho, A. Rius, M.P. Callao, M.S. Larrechi, *Anal. Chim. Acta* **445** (2001) 213.
- [16] J.S. Shenk, M.O. Westerhaus, *U.S. Patent 4866644*. Sept. 12 (1989).
- [17] E. Bouveresse, D.L. Massart, P. Dardenne, *Anal. Chim. Acta* **297** (1994) 405.
- [18] M. Forina, G. Drava, C. Armanino, R. Boggia, S. Lanteri, R. Leardi, P. Corti, P. Conti, R. Giangiacomo, C. Galliena, R. Bigoni, I. Quartari, C. Serra, D. Ferri, O. Leoni, L. Lazzeri, *Chemom. Intell. Lab. Syst.* **27** (1995) 189.
- [19] B. Walczak, E. Bouveresse, D.L. Massart, *Chemom. Intell. Lab. Syst.* **36** (1997) 41.
- [20] C. Tan, M. Li, *Anal. Sci.* **23** (2007) 201.
- [21] R. Goodacre, É.M. Timmins, A. Jones, D.B. Kell, J. Maddock, M.L. Heginbothom, J.T. Magee, *Anal. Chim. Acta* **348** (1997) 511.
- [22] L. Duponchel, C. Ruckebusch, J.P. Huvenne, P. Legrand, *J. Mol. Struct.* **480** (1999) 551.

- [23] N.A. Woody, R.N. Feudale, A.J. Myles, S.D. Brown, *Anal. Chem.* **76** (2004) 2595.
- [24] W. Ni, S.D. Brown, R. Man, *Anal. Chim. Acta* **661** (2010) 133.
- [25] G. Siano, Goicochea H., *Chemom. Intell. Lab. Syst.* **88** (2007) 204.
- [26] T.M. Alam, M.K. Alam, *J. Chemom.* **24** (2010) 261.
- [27] Z.P. Chen, L.M. Li, R.Q. Yu, D. Littlejohn, A. Nordon, J. Morris, A. Dann, P.A. Jeffkins, M.D. Richardson, S.L. Stimpson, *Analyst* **136** (2011) 98.
- [28] A.J. Myles, T.A. Zimmerman, S.D. Brown, *Appl. Spectrosc.* **60** (2006) 1198.
- [29] F.W. Koehler IV, G.W. Small, R.J. Combs, R.B. Knapp, R.T. Kroutil, *Anal. Chem.* **72** (2000) 1690.
- [30] S.A. Roussel, C.L. Hardy, C.R. Hurburgh, G.R. Rippke, *Appl. Spectrosc.* **55** (2001) 1425.
- [31] Y. Xu, R.G. Brereton, *Anal. Bioanal. Chem.* **388** (2007) 655.
- [32] M. Padilla, A. Perera, I. Montoliu, A. Chaudry, K. Persaud, S. Marco, *Chemom. Intell. Lab. Syst.* **100** (2010) 28.
- [33] C.V. Di Anibal, M. Odena, I. Ruisánchez, M.P. Callao, *Talanta* **79** (2009) 887.
- [34] L.A. Stone, R.W. Kennard, *Technometrics* **11** (1969) 137.
- [35] B.R. Kowalski, *Chemometrics: Mathematics and statistics in chemistry*, D. Reidel publishing company, Dordrecht, Holland (1984).
- [36] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Koch, "PLS Toolbox Version 3.5 for use with Matlab™", Eigenvector Research, Inc. Manson, WA, USA.

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

C H A P T E R 4

C Conclusions

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

This chapter summarizes the general conclusions of the work presented in this thesis because the conclusions concerning each individual study have been drawn at the end of each paper in Chapter 3.

This thesis deals with screening adulterated culinary spices to determine the presence of Sudan red dyes. Nowadays, despite regulatory surveillance, the use of non-permitted dyes in loose or non-branded spices that are mainly found in rural markets still continues to be a food safety problem in some countries. Therefore, simple and reliable methods for being used as monitoring tools need to be developed when such situation is found. In particular, these methods are less-time consuming than the usual methods used to detect such dyes in foods.

The methodologies presented in this thesis are based on spectroscopic techniques combined with multivariate analysis and represent feasible methods for carrying out routine analyses or specific controls requiring a rapid response. They can be extrapolated to other food issues such as adulteration, fraud, etc. and can be implemented by using other analytical techniques, instrumental modalities and classification techniques.

More specifically, the following conclusions can be drawn regarding the pre-defined objectives that were set out in Chapter 1:

1. Study of spectrometric techniques such as UV-Visible, ¹H-NMR and Raman to obtain the multivariate data (spectra).

- The working conditions for UV-Visible, ¹H-NMR and Raman techniques have fine tuned in order to determine the optimal experimental parameters for obtaining the spectra that are subjected to multivariate analysis.

2. Establishment and application of chemometric tools such as exploratory data analysis, supervised classification techniques, data processing and variable selection techniques to extract the maximum possible information from the spectral data.

- The exploratory analysis is a useful way of giving an initial view of the natural groupings formed by the different classes of samples studied.

- The classification models developed in this thesis makes it possible to predict whether a certain sample is safe for consumption or whether it is adulterated with Sudan dyes. If the sample is adulterated, the models can also identify the specific dye contained in this sample.

- The data processing and variable selection methods enable more accurate and summarized spectral information to be obtained. Data processing has been shown to be necessary in Raman spectroscopy in order to correct spectral shifts and interferences. The variable selection methods are essential for selecting the most important variables from high-dimensional data such as NMR, because they improve the performance of the multivariate model. There is no single best variable selection method; the choice of method depends on the nature of the problem.

3. Implementation of data fusion to improve the classification of results derived from the individual spectrometric techniques.

A data fusion strategy allows the information provided by different sources to be brought together. In this regard, it offers a way of enhancing and strengthening the information relating to a specific analytical problem. Moreover, different data fusion strategies can be used which makes them a versatile tool that can be applied to different spectral data.

4. Study of multivariate transfer techniques (standardization) to maintain the original classification model through a lapse of time.

Standardization methods provide a useful way of resolving situations where the experimental measurements have been taken under different conditions in which non-controlled changes have caused differences in the recorded spectra. In this regard, the ability of a classification model can be maintained over time by implementing simple strategies with associated time saving and cost benefits. This is particularly interesting in situation where the samples needed to build the model are expensive, numerous, hazardous or difficult to treat experimentally.

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

A p p e n d i x

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

A.1 List of papers presented in this thesis

1. Carolina V. Di Anibal, Marta Odena, Itziar Ruisánchez, M. Pilar Callao, *Determining the adulteration of spices with Sudan I, II, III and IV by UV-Visible spectroscopy and multivariate classification techniques*, *Talanta* 79 (2009) 887-892.
2. Carolina V. Di Anibal, Itziar Ruisánchez, M. Pilar Callao, *High-Resolution ^1H -Nuclear Magnetic Resonance spectrometry combined with chemometric treatment to identify adulteration of culinary spices with Sudan dyes*, *Food Chemistry* 124 (2011) 1139-1145.
3. Carolina V. Di Anibal, Lluís F. Marsal, M. Pilar Callao, Itziar Ruisánchez, *Surface Enhanced Raman Spectroscopy (SERS) and multivariate analysis as a screening tool for detecting Sudan I dye in culinary spices*, Submitted for publication (2011).
4. Carolina V. Di Anibal, M. Pilar Callao, Itziar Ruisánchez, *^1H -NMR variable selection approaches for classification. A case study: the determination of adulterated foodstuffs*, *Talanta* 86 (2011)316-323.
5. Carolina V. Di Anibal, M. Pilar Callao, Itziar Ruisánchez, *^1H -NMR and UV-Visible data fusion for determining Sudan dyes in culinary spices*, *Talanta* 84 (2011) 829-833.
6. Carolina V. Di Anibal, Itziar Ruisánchez, Mailén Fernández, Rafel Forteza, Victor Cerdà, M. Pilar Callao, *Standardization of UV-Visible data in a food adulteration classification problem*, Submitted for publication (2011).

A.2 Meeting Contributions

Classification of Sudan dyes with rapid methodologies based on spectroscopic techniques and the chemometric approach

Carolina V. Di Anibal, M. Pilar Callao and Itziar Ruisanchez

11th International Conference on Chemometrics for Analytical Chemistry (CAC),
Montpellier, France, (2008)

Poster communication

Determination of food adulteration by classification techniques

Carolina V. Di Anibal, M. Pilar Callao and Itziar Ruisanchez.

III Workshop de Quimiometría, Burgos, Spain, (2008)

Poster with flash communication

Determination of the paprika spice adulteration with Sudan dyes (I to IV) by means of chemometric techniques

Carolina V. Di Anibal, M. Pilar Callao and Itziar Ruisanchez

12th Conference on Instrumental Analysis (JAI), Barcelona, Spain (2008)

Poster communication

Determination of foodstuffs adulteration with NMR spectrometric technique and chemometric approach

Carolina V. Di Anibal, M. Pilar Callao and Itziar Ruisanchez

XV Reunión de la Sociedad Español de Química Analítica, San Sebastián, Spain
(2009)

Poster communication

¹H-NMR and UV-Visible spectra fusion for determining Sudan dyes in culinary spices

Carolina V. Di Anibal, M. Pilar Callao and Itziar Ruisanchez

VII Colloquium Chemometricum Mediterraneum, Granada, Spain (2010)

Oral communication

Surface Enhanced Raman Spectroscopy (SERS) and multivariate analysis as a screening tool for detecting Sudan I dye in culinary spices

Carolina V. Di Anibal, M. Pilar Callao and Itziar Ruisanchez

13th Conference on Instrumental Analysis (JAI), Barcelona, Spain (2011)

Poster communication

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011

UNIVERSITAT ROVIRA I VIRGILI

DETERMINATION OF BANNED SUDAN DYES IN CULINARY SPICES THROUGH SPECTROSCOPIC TECHNIQUES
AND MULTIVARIATE ANALYSIS

Carolina Vanesa Di Anibal

DL:T-1809-2011